Konstruktion eines Situational Judgment Tests für die Führungsdiagnostik auf der Grundlage des Act Frequency Approachs und des Wertequadrats

Abhandlung
zur Erlangung der Doktorwürde
der Philosophischen Fakultät
der Universität Zürich

vorgelegt von
Patrick Boss
von Sigriswil / BE

Angenommen im Herbstsemester 2010 auf Antrag von
Herrn Prof. Dr. Klaus Jonas und Herrn Prof. Dr. François Stoll

Zürich, 2012

Abstract

Für die standardisierte Beurteilung der mittels Literaturstudium und der Critical Incident Technique für unteres Milizkader der Schweizer Armee bestimmten Führungskompetenzen Durchsetzungsfähigkeit, Kontaktfähigkeit und Verantwortungsbewusstsein wird ein Persönlichkeits-Fragebogen entwickelt, welcher auf den computergestützten Testanlagen in den sechs Rekrutierungszentren als Screening-Instrument zum Einsatz gelangt.

Das Ausgangsmaterial für die Formulierung der Situationen (Item-Stämme) und der Verhaltensalternativen der Items des in Form eines Situational Judgment Tests umgesetzten Persönlichkeits-Fragebogens bildet ein Pool von anhand des Act Frequency Approachs gesammelten prototypischen Verhaltensweisen für jede der drei Kompetenzen. Als Konstruktionsrational für die Formulierung der jeweils vier Verhaltensalternativen pro Situation wird das Wertequadrat (Helwig, 1948) eingesetzt.

Die erste Version des Fragebogens umfasst pro Kompetenzdimension je 13 Items im Forced-Choice-Format und weist Skalenreliabilitäten zwischen α = .56 und .71 auf. Um die Reliabilitäten zu steigern, wird in der zweiten Version ein vierstufiges, likert-skaliertes Antwortformat eingesetzt, wobei jede der vier Verhaltensalternativen pro Item einzustufen ist. Dies führt bei einer Testkürzung auf zehn Items pro Skala zu Reliabilitäten zwischen α = .81 und .93. Explorative Faktorenanalysen und der Vergleich mit zwei Persönlichkeitsfragebogen belegen die Konstruktvalidität des Fragebogens. Zudem zeigt sich, dass der Bekanntheitsgrad der Item-Stämme mit der Trennschärfe der Items zusammenhängt, was die Wichtigkeit der zielgruppenspezifischen Konstruktion von Situational Judgment Tests belegt.

Ein wichtiges Teilziel dieser Entwicklung stellt die Akzeptanz des Testverfahrens bei den Stellungspflichtigen dar. In mehreren Studien wird aufgezeigt, dass das Hinzufügen einer die geschilderte Situation illustrierenden Fotografie wesentlich zur Erhöhung der Akzeptanz des Verfahrens beiträgt und diese insgesamt deutlich höher ausfällt, als bei einem herkömmlichen Persönlichkeits-Fragebogen.

Schlagworte: Persönlichkeitsfragebogen, Führungsdiagnostik, Schweizer Armee, Rekrutierung, Situational Judgment Test, Anforderungsprofil für unteres Milizkader, Critical Incident Technique, Act Frequency Approach, Wertequadrat, Akzeptanz

Abstract

This dissertation develops a personality questionnaire for standardized assessment of leadership qualities/skills determined important (based on a review of the literature and using the critical incident technique) for lower-level militia cadre in the Swiss Armed Forces: assertiveness, interpersonal skills, and sense of responsibility. The questionnaire will be used as a screening instrument at computer-supported testing facilities at Switzerland's six recruitment centers.

The personality questionnaire is designed as a situational judgment test. For development of the situations (item stems) and the behavior options (item responses) in the questionnaire, the raw material is a pool of prototypical behaviors for each of the three leadership qualities/skills, compiled using the act frequency approach. The construction rationale for the formulation of the four behavior options per situation is Helwig's (1948) square of values.

The first version of the questionnaire has 13 items in a forced choice format per leadership dimension, and the reliability of the scales ranges from α = .56 to .71. To increase the reliability, in the second version of the questionnaire a 4-point Likert response scale is used; respondents grade each of the four behavior options per item using the scale. When the test is shortened to 10 items per scale, this results in reliabilities ranging from α = .81 to .93. Explorative factor analyses and comparison with two personality questionnaires give evidence of the construct validity of the questionnaire. In addition, the level of familiarity of the item stems is associated with the discrimination power of the items, which demonstrates the importance of target group-specific construction of situational judgment tests.

An important sub-goal of this development is acceptance of the test by the military conscripts. Several studies have demonstrated that adding photographs that illustrate the situations described in the items contributes significantly to increased acceptance of a test, with acceptance being clearly higher than acceptance of conventional personality questionnaires.

Keywords: Personality questionnaire, leadership diagnostics, Swiss Armed Forces, recruitment, situational judgment test, requirements for lower-level militia cadre, critical incident technique, act frequency approach, Helwig's square of values, acceptance

Danksagung

Die umfangreichen Arbeiten, welche im Zusammenhang mit der hier beschriebenen Testkonstruktion erfolgten, wären ohne die Mithilfe von Studierenden und meinem Team unmöglich gewesen. Ihnen gilt mein besonderer Dank:

Meinen Mitarbeiterinnen Andrea Boss-Skupnjak, Klara Jerabek, Eva Jöri, Irène Kunz-Betschart, Esther Maier, Katrin Roduner, Vera Schiess-Maier, Andrea Schnyder und Sibylle Wirth.

Den Studentinnen und Studenten Raphael Bauhofer, Dieter Bösser, Astrid Bühler Ruedin, Erika Deiss, Andrea Emerson, Crista Henggeler, Andrea Imper, Vera Maier, Fabienne Moroge, Marianne Schibli und Tanja Selk.

Prof. Dr. Klaus Jonas danke ich für das kritische Durchlesen des Manuskripts, die wertvollen Hinweise und nicht zuletzt auch für seine Geduld.

Prof. em. Dr. François Stoll danke ich für sein Vertrauen, welches er mir in meiner Funktion als Projektleiter während vieler Jahre entgegengebracht hatte und dass er auch lange nach seinem Rücktritt noch bereit ist, die Arbeit eines ehemaligen Studenten und Assistenten zu begutachten.

Der Auftrag vom Bund, den psychologischen Teil der Rekrutierung zu konzipieren und umzusetzen war für mich eine einmalige Chance, welche mir viele spannende Momente, viele Erfahrungen und ein umfangreiches Wissen bescherte. Mein Dank gilt den Herren Div a D Waldemar Eymann, KKdt Dominique Andrey, Oberst i Gst a D Willi Staubli, Br Philippe Rebord (Et votre thèse? – Enfin terminée!) und nicht zuletzt auch Oberst i Gst René Baumann, mein „Kamerad der ersten Stunde".

Dr. Chris Roetheli danke ich für den gedanklichen Austausch, seine kameradschaftliche Unterstützung und seinen Glauben an meine Dissertation.

Ganz herzlich möchte ich meinen Eltern für ihre Unterstützung während meines Studiums danken und meiner Frau, welche mich bis zuletzt bedingungslos in meinen Dissertationsplänen unterstützt hat.

Patrick Boss, September 2010

Inhaltsverzeichnis

1. Ausgangslage und Zielsetzung

1.1 Von der Aushebung zur Rekrutierung der Schweizer Armee

Im Zusammenhang mit der Armeereform „Armee XXI" der Schweizer Armee, welche 2004 umgesetzt wurde, begannen die Armeeplaner 1999 mit der Neukonzeption der Aushebung. Sie entschieden, dass die Kantone von der Durchführung der seit Generationen dezentral im ganzen Land stattfindenden Aushebung entbunden werden. Die neu zu schaffende Rekrutierung ist nun eine Angelegenheit des Bundes, dauert zwei bis drei Tage und fand ursprünglich in sieben, heute noch in sechs Rekrutierungszentren statt. Dabei untersuchen Spezialisten die körperliche und neu auch die psychische Verfassung des Stellungspflichtigen sehr ausführlich und mit modernen Untersuchungsmethoden. Dies stellt einen einschneidenden Paradigmenwechsel in der Beurteilung der Diensttauglichkeit dar, da zuvor ein zweistufiges Verfahren zum Einsatz gelang: An der Aushebung führten die Ärzte eine Grobtriage durch, bei welcher sie nur Stellungspflichtige mit klaren und deutlich ausgeprägten Symptomen militärdienstleistungsrelevanter Störungen oder Beeinträchtigungen als untauglich erklärten. Die psychologische Untersuchung beschränkte sich dabei auf die Durchführung eines Intelligenztests. Eine zweite Tauglichkeitsbeurteilung fand dann in den ersten Wochen der Rekrutenschule statt, was dazu führte, dass nochmals 15 – 20% des Einrückungsbestandes entlassen und zu einem grossen Teil auch vom Militärdienst befreit wurde (Boss, Vetter, Frey & Lupi, 2003; Frey, Huber & Lupi, 2003). Dieses Vorgehen bedingte, dass in den Schulen mehr Ausbildungskapazitäten zur Verfügung standen, als schlussendlich tatsächlich benötigt wurden. Zudem kostet ein Rekrut rund 200.– Franken pro Diensttag an Unterkunft und Verpflegung. Dieses bei Rekrutenschul-Abbrechern unnötig ausgegebene Geld wollte man mit einer detaillierteren Vorselektion einsparen respektive genau dafür einsetzen.

Die Projektleitung formulierte folgende Vorgaben für die neue Rekrutierung (Schweizerische Armee, 2000):

- Die Anzahl Abbrüche der Rekrutenschule soll minimiert werden. Eine effizientere Rekrutierung soll den oft teuren und zeitintensiven Fehlbeurteilungen wirksam entgegentreten.
- Die Eignungsprüfungen im Hinblick auf Diensttauglichkeit und Funktionszuteilung müssen neu konzipiert werden. Sie haben dabei aktuellen und auch zukünftigen Anforderungen zu genügen.

- Eine erste Erfassung des Kaderpotentials soll anlässlich der Rekrutierung erfolgen.
- Die Eignungsprüfung soll wegen der Gültigkeit der Daten in der Regel ca. drei bis sechs (maximal zwölf) Monate vor der Grundausbildung erfolgen.
- Der Prüfungsablauf soll individuell auf die vorgesehene Funktion ausgerichtet werden und soll ein bis drei Tage dauern.

Ende 1999 haben die Projektverantwortlichen auf militärischer Seite der Abteilung Angewandte Psychologie der Universität Zürich den Auftrag erteilt, im Rahmen der Konzeptionsstudie Rekrutierung ein Konzept zu erstellen, wie der psychologische Teil der Rekrutierung neu und aussagekräftiger als bisher gestaltet werden kann, wobei psychische Ressourcen, Leistungsaspekte, Persönlichkeit, soziale Kompetenzen und Interessen abgeklärt werden sollen. Diese Abklärungen haben die Bereiche Grundrekrutierung, Kaderbeurteilung und die Beurteilung der höheren Kader zu umfassen. In meiner Funktion als universitätsseitiger Projektleiter des psychologischen Teils der Rekrutierung formulierte ich folgende Ziele, welche mit der neuen Rekrutierung zu erreichen sind (Stoll, Boss & de With, 2000):

1. Senkung des Anteils an Nichtausexerzierten (Rekrutenschul-Abbrecher) durch eine genaue Abklärung der psychischen Ressourcen.
2. Für die Zuteilung zu den Funktionen sollen aussagekräftigere Informationen über den Stellungspflichtigen vorliegen.
3. Beurteilung des Kaderpotenzials anhand einiger relevanter Persönlichkeitseigenschaften und Fähigkeiten und bessere Verteilung von Rekruten mit Kaderpotenzial über die verschiedenen Truppengattungen.
4. Mit einer ausführlichen führungsbezogenen Eignungsabklärung soll die Entscheidungsfindung bezüglich der Zulassung an Kaderschulen unterstützt werden.
5. Die vielfältigen psychologischen Abklärungen, wie zum Beispiel Kaderbeurteilungen oder Eignungsabklärungen für Auslandeinsätze, sollen in ein Gesamtkonzept eingebaut werden, damit Doppelspurigkeiten vermieden und Synergien genützt werden können.

Die für die Erreichung der zwei wichtigsten Ziele der neuen Rekrutierung – der umfassenden Tauglichkeitsabklärung und der möglichst optimalen Passung zwischen den Fähigkeiten des Stellungspflichtigen und den Anforderungen der militärischen Funktion, welcher ihn der Rekrutierungsoffizier zuteilt – lassen sich

mit dem Satz „den Stellungspflichtigen gründlich kennen lernen“ (Boss & Baumann, 2003) umschreiben. Da die Rekrutierung der erste Kontakt der jungen Männer mit dem Militär darstellt und dabei die Weichen für die folgenden zehn Militärdienst-Jahre gestellt werden, ist es zudem absolut zentral, dass der Stellungspflichtige den Eindruck eines professionellen und fairen Verfahrens erhält, in welchem er als eigenständige Person ernst genommen wird. Wenn er sich als Nummer, als einer unter vielen behandelt fühlt oder den Eindruck gewinnt, dass seine Leistung im Stangenklettern seine militärische Laufbahn bestimmt, wird dies jedoch kaum der Fall sein. Zudem sind die Zeiten, in welchen Jugendliche akzeptierten, dass sie in einem schlecht beleuchteten Theoriesaal in einer Zivilschutzanlage etlichen "Papierkram" zu erledigen haben, endgültig vorbei.

Dass schon anlässlich der Rekrutierung eine Abklärung des Kaderpotenzials durchzuführen ist, war ein weiterer, nicht unumstrittener Paradigmenwechsel bei der Auswahl und Ausbildung des militärischen Kaders. Folgende zwei Ziele standen dabei im Mittelpunkt (Stoll et al., 2000, S. 32):

1. Der Rekrutierungsoffizier teilt Stellungspflichtige mit Kaderpotenzial gleichmässig den verschiedenen Truppengattungen und Funktionen zu. Damit ist gewährleistet, dass in allen Bereichen der Armee ein genügend grosser Anteil an fähigem Kadernachwuchs vorhanden ist.
2. Den mit der Kaderselektion in den Rekrutenschulen beauftragen Berufskadern stellt das Rekrutierungszentrum Informationen über die Kadereignung ihrer Rekruten zur Verfügung. Büttiker und Stoller (1989) zeigten in ihrer Untersuchung auf, dass charakterliche Schwächen die am meisten genannten Gründe dafür sind, dass man einen Anwärter nicht zur Weiterausbildung vorschlägt. Da in der Armee XXI schon sehr früh entschieden werden muss, wer eine Kaderlaufbahn einschlagen wird, ist die Einschätzung des Charakters der Rekruten nur grob möglich. Testresultate können hier als wichtige, zusätzliche Informationsquelle dienen.

Es stellte sich hier die Frage, ob es denn schon möglich sei, bei einem 19jährigen das Kaderpotenzial zu erfassen. Dabei ist davon auszugehen, dass die Persönlichkeitsentwicklung in diesem Alter noch nicht vollständig abgeschlossen ist, zumal diese einen lebenslangen Prozess darstellt. Mit der Festigung des Selbstkonzeptes in der Kindheit und der Jugend stabilisieren sich jedoch auch die Persönlichkeitseigenschaften auf ein im Erwachsenenalter über viele Jahre konstantes Niveau (Asendorpf, 1999). Somit erscheint es aus einer entwicklungspsychologischen Sicht auch vertretbar zu sein, bei 19jährigen das Kaderpotenzial zu erfassen, da deren Persönlichkeit schon so weit gefestigt ist, dass nicht mehr

mit tief greifenden Veränderungen zu rechnen ist. Weit stabiler als die absolute Ausprägung einer Persönlichkeitseigenschaft ist jedoch die relative im Vergleich zu anderen Personen: So wird zum Beispiel der erfolgorientierteste einer Gruppe Jugendlicher auch zehn Jahre später mit grosser Wahrscheinlichkeit der oder zuminderst einer der Erfolgorientiertesten dieser Gruppe sein (z. B. Conley, 1984). Auf die Situation in den Rekrutierungszentren umgesetzt bedeutet dies, dass wenn der Rekrutierungsoffizier aus einer Gruppe von Stellungspflichtigen diejenigen mit den ausgeprägtesten Kadereigenschaften auswählt, diese auch zehn Jahre später zur Gruppe mit guten Führungsqualitäten gehören.

Das Problem im Zusammenhang mit der Erfassung des Kaderpotenzials anlässlich der Rekrutierung stellt sich viel grundlegender: Ist es möglich, quasi in einem Laborsetting – der Rekrutierung – ohne Vorkenntnisse über den Stellungspflichtigen und ohne ausreichende Beobachtungsmöglichkeiten anhand von einigen wenigen Angaben zur kognitiven Leistungsfähigkeit ergänzt durch reine Selbstbeschreibungen eine valide Aussage zum Kaderpotenzial zu machen? Und wenn ja, wie lange hat diese Voraussage Gültigkeit? Letztere Frage lässt sich in unserem Fall leicht beantworten, da die Rekrutierung ungefähr sechs bis zwölf Monate vor dem Eintritt in die Rekrutenschule stattfindet und dort die zukünftigen Kader bereits nach sieben Wochen Grundausbildung in die Kaderschule wechseln. Die in der Rekrutierung durchgeführte Beurteilung muss also ungefähr ein Jahr Gültigkeit haben, bis diese durch eine erneute Beurteilung überprüft wird. Somit ist die zentrale Frage, ob die im Rahmen einer Kaderselektion eingesetzten Testverfahren überhaupt zu validen Ergebnissen führen. Schmidt und Hunter (1998a, 1998b) konnten in ihrer Meta-Analyse nachweisen, dass Arbeitsproben, strukturierte Interviews und Intelligenztests die validesten Prädiktoren von Berufsleistung sind. Die Meta-Analyse von Bertua, Anderson und Salgado (2005) zeigt zudem auf, dass die Komplexität der Arbeitsinhalte den Zusammenhang zwischen der Leistung im Intelligenztest und der Berufsleistung moderiert, so dass eine Intelligenztestung bei anspruchsvolleren Tätigkeiten zu valideren Vorhersagen führt als bei einfachen. Diese – und eine Vielzahl hier nicht aufgeführter – Forschungsergebnisse belegen, dass Intelligenztests valide Prädiktoren in der Kaderselektion darstellen.

Umstritten ist hingegen der Einsatz von Persönlichkeits-Fragebogen im Rahmen der Personalselektion. Auch wenn die im deutschen Sprachraum in den siebziger Jahren geführte Grundsatzdiskussion über die Zulässigkeit deren Einsatzes heute überwunden ist (z. B. Grubitzsch & Rexilius, 1978; Pulver, Lang & Schmid, 1978; Schweizerische Gesellschaft für Psychologie, 1975, 1976. Ich werde darauf noch ausführlich in Kapitel 5.1 eingehen), gehen auch heute noch die Meinungen der Fachexperten über deren Nutzen auseinander. Die Zeitschrift

Human Performance hat 2005 eine Ausgabe diesem Thema gewidmet und Experten auf dem Gebiet der Personalselektion zu Wort kommen lassen (Barrick & Mount, 2005; Hogan, 2005; Hough & Oswald, 2005; Murphy & Dzieweczynski, 2005; Ones, Viswesvaran & Dilchert, 2005). Murphy und Dzieweczynski führen an, dass viele der Probleme, welche Guion und Gottier 1965 zum Einsatz von Persönlichkeits-Fragebogen in der Personalselektion nannten, bis heute nicht gelöst sind. Währenddem Intelligenztests zu allen Tätigkeiten, bei deren Ausführung kognitive Prozesse involviert sind, einen Zusammenhang aufweisen, müssen die Dimensionen eines Persönlichkeits-Fragebogens gut auf die Inhalte der Tätigkeit abgestimmt sein, um eine befriedigende Übereinstimmung zu erzielen. Dabei ergeben sich Schwierigkeiten bei der Bestimmung der für die Ausübung einer Tätigkeit erforderten Persönlichkeitseigenschaften, welche sich auch nicht mit ausgeklügelten Methoden, wie zum Beispiel der von Tett und Burnett (2003), zufrieden stellend lösen lassen. So seien die Validitäten von Persönlichkeits-Fragebogen immer noch zu tief, um von einem Mehrwert deren Einsatzes sprechen zu können. Dieser Einschätzung widersprechen Hough und Oswald (2005) mit dem Verweis auf die Meta-Analyse von Hough und Furnham (2003), welche unter anderem aufzeigt, dass der Zusammenhang zwischen Persönlichkeitseigenschaften und der Berufsleistung auch vom Tätigkeitsspektrum des jeweiligen Berufs abhängt. Zudem führen sie auf, dass bei komplexen Kriterien die Korrelation mit umfassenden, breiten Persönlichkeitsmerkmalen höher ist als mit eng umschriebenen, homogenen. Hough und Oswald verweisen auch auf die Bedeutung der Situation als Moderator der Kriteriumsvalidität: Beaty, Cleveland und Murphy (2001) konnten aufzeigen, dass in „starken Situationen" (Arbeitssituationen, in welchen klare Hinweise bestehen, wie man sich zu verhalten hat) im Gegensatz zu „weichen Situationen" (Arbeitssituationen mit uneindeutigen Hinweisen) die Korrelationen zur Leistungsbeurteilung durch den Vorgesetzten tiefer ausfällt (r = .13 vs. .29).

Auch Barrick und Mount (2005) sind überzeugt, dass Persönlichkeitseigenschaften bei der Berufsausübung eine wichtige Rolle spielen, wobei sie vor allem auf die Bedeutung des Situationskontextes hinweisen:

> It is now apparent that we must consider situational demands that extend beyond the immediate demands of the job to fully define whether the context is relevant to a particular personality trait. We believe personality will have its greatest effect on behavior when the situation, broadly defined by the demands of the job, group, and organization, is relevant to the trait's expression and is weak enough to allow the person to choose how to behave in that situation. ... To assure the more effective use of personality measures, we must carefully consider the need to aggregate

> behavior. Such aggregation should be guided by the need to balance our ability to predict consistency in behavior without eliminating or "factoring out" the influence of the situation. Practical and theoretical understanding will only occur if we account for the influence of the person and the situation on behavior. (S. 368-369)

Morgeson, Campion, Dipboye, Hollenbeck, Murphy und Schmitt haben mit ihrem 2007 veröffentlichten Beitrag „Reconsidering the use of personality tests in personnel selection contexts" eine erneute Diskussion über die Zulässigkeit des Einsatzes von Persönlichkeits-Fragebogen in der Personalselektion ausgelöst (Repliken auf diesen Artikel: Morgeson, Campion, Dipboye, Hollenbeck, Murphy & Schmitt, 2007b; Ones, Dilchert, Viswesvaran & Judge, 2007; Tett & Christiansen; 2007). Sie fokussierten dabei auf drei, ihrer Meinung nach noch immer ungelöste Probleme: Die bewusste Antwortverfälschung (*Faking,* siehe dazu z. B. Birkeland, Manson, Kisamore, Brannick & Smith, 2006), die geringe prädiktive Validität zu Berufsleistung und die zum Teil ungeeigneten Inhalte einiger Verfahren. Das wichtigste Argument gegen den Einsatz von Persönlichkeits-Fragebogen in der Personalselektion stellt dabei die prädiktive Validität dar, welche zum Beispiel bei den Big Five zwischen r = -.02 und .15 liegt, wobei Gewissenhaftigkeit die höchsten Werte erzielt (Barrick & Mount, 1991; Hurtz & Donovan, 2000; Salgado, 1997). Tett und Christiansen (2007) und Ones et al. (2007) wiesen in ihrer Replik unter anderem auf die sehr umfangreiche Meta-Analyse von Judge, Bono, Ilies und Gerhardt (2002) hin, welche Zusammenhänge zwischen den Big Five und Führungsfähigkeit (*Leadership*) von r = .06 bis .22 (ρ = .08 bis .31; R = .30 bis .45) aufzeigen konnten. Tett und Christiansen machen weiter darauf aufmerksam, dass eng gefasste Persönlichkeitsdimensionen eine bessere Vorhersage von Berufsleistung erlauben als die breiten Dimensionen des Big Five-Modells (Rothstein & Goffin, 2006), was jedoch im Widerspruch zur oben aufgeführten Aussage von Hough und Furnham (2003) steht. Ones et al. zeigen anhand einer Aufstellung, welche mehrere Studien umfasst, dass die Zusammenhänge zwischen den einzelnen Big Five-Dimensionen und Berufsleistung in Abhängigkeit von der Berufsgattung unterschiedlich hoch ausfallen. Zudem weisen sie auf Studien zur inkrementellen Validität von Persönlichkeitsdaten gegenüber allgemeiner Intelligenz hin, welche zwischen .07 und .16 beträgt (z. B. Ones & Viswesvaran, 2001). Sie schliessen daraus, dass „there is incremental validity to be gained from using personality measures in predicting overall job performance, even when cognitive variables are already included in a given selection system" (Ones et al., 2007, S. 1010).

Nicht ganz einig sind sich Morgeson et al. (2007a) über die Auswirkungen des Fakings. So sollen diese auf die Validität nur gering sein, da alle Bewerber in

etwa in demselben Ausmass ihre Angaben verfälschen. Zudem sei das bewusste Steuern der Antworten eine Kompetenz, welche auch im Berufsalltag von Bedeutung ist, da man auch dort Situationen analysieren muss, um ein möglichst Erfolg versprechendes Verhalten zu zeigen (z. B. Marcus, 2003). Diesen Ansichten widersprechen Tett und Christiansen mit Nachdruck, da erwiesen ist, dass die Bewerber ihre Antworten in einem Persönlichkeits-Fragebogen unterschiedlich stark verfälschen (z. B. Tett, Anderson, Ho, Yang, Huang & Hanvongse, 2006) und dies einen Einfluss auf die Validität des Verfahrens hat (z. B. Griffith, Chmielowski & Yoshita, 2007) und darauf, wer schlussendlich eingestellt wird (Rossé, Stecher, Miller & Levin, 1998). Tett und Christiansen (2007, S. 984) kommen auf Grund ihres Literaturstudiums zum Schluss, dass „applicant faking attenuates personality test validity but enough trait variance remains to be useful for predicting job performance."

Will man Persönlichkeits-Fragebogen trotz des Mangels an (ausreichend) hoher prädiktiver Validität einsetzen, so machen Morgeson et al. (2007a) darauf aufmerksam, dass „... customized personality measures that are clearly job-related in face valid ways might be more easily explained to both candidates and organizations" (S. 721). Tett und Christiansen (2007) weisen jedoch darauf hin, dass Bewerber in der Regel mehr und erfolgreicher faken, wenn ihnen bekannt ist, welche Dimensionen der Test erfasst, oder wenn die Items in Bezug auf das zugrunde liegende Konstrukt sehr transparent sind (z. B. Alliger, Lilienfeld & Mitchell, 1995). Ones et al. raten zudem vom Einsatz des Forced-Choice-Antwortformates zur Reduzierung von Faking ab, weil die daraus resultierenden ipsativen Daten zu psychometrischen Problemen führen, zum Beispiel bei der Berechnung der Reliabilitäten oder der Durchführung von Faktorenanalysen bei der Bestimmung der Konstruktvalidität (z. B. Dunlap & Cornwell, 1994; Hicks, 1970; Johnson, Wood & Blinkhorn, 1988; Meade, 2004; Tenopyr, 1988).

Diese Ausführungen zusammenfassend lässt sich festhalten, dass nicht grundsätzlich vom Einsatz von Persönlichkeits-Fragebogen in der Personalselektion abgeraten werden kann, dass man jedoch deren Möglichkeiten und Grenzen genau kennen muss. Sehr wichtig scheint dabei – im Gegensatz zum Einsatz von Intelligenztests – zu sein, dass die im Fragebogen erfassten Persönlichkeitsdimensionen auf die Anforderungen der zu besetzenden Arbeitsstelle abgestimmt sind.

Im Hinblick auf den Einsatz eines Persönlichkeits-Fragebogens im Rahmen der Kaderbeurteilung in den Rekrutierungszentren sind die Ergebnisse aus der Studie von Mueller-Hanson, Heggestad und Thornton (2003) sehr bedeutsam. Diese konnten in einem Experiment nachweisen, dass die Wahl der Selektions-

strategie – Besten-Selektion (*select in*) im Vergleich zur Negativ-Selektion (*select out*) – eine Auswirkung auf die Validität von Persönlichkeits-Fragebogen hat. Sie wiesen 444 Studenten zufällig einer Experimental- und einer Kontrollgruppe zu, wobei alle die englischsprachige Version des Leistungsmotivations-Inventar (LMI) auszufüllen hatten. Der Kontrollgruppe teilten sie mit, dass sie den Fragebogen ehrlich ausfüllen sollen und es sehr wichtig sei, dass sie sich so beschreiben, wie sie wirklich sind. Die Experimentalgruppe instruierten sie dahingehend, dass für den zweiten Teil der Studie Personen gesucht würden, welche motiviert und gewissenhaft sind und hart arbeiten, wozu sie den LMI als Selektionsinstrument einsetzten. Dabei hätten nur diejenige, welche auch am zweiten Teil der Studie teilnehmen dürfen, Aussicht auf den Gewinn eines Sofortpreises von 20 Dollar. Als zweiten Test legten Mueller-Hanson et al. allen Studienteilnehmern einen eigens dafür entwickelten Leistungstest vor, welcher aus 50 einfachen, aber zeitaufwändigen und ermüdenden Items bestand. Um den Zusammenhang zur Leistungsmotivation zu verstärken, instruierten sie die Teilnehmer, dass es ihnen freigestellt ist, wie lange sie diese Aufgaben bearbeiten.

Die Studenten der Kontrollgruppe erzielten durchschnittlich 214.70 Punkte im LMI und 40.50 Punkte im Leistungstest, diejenigen der Experimentalgruppe 225.26 respektive 40.08 Punkte. Die Korrelation zwischen dem LMI und dem Leistungstest beträgt bei der Kontrollgruppe r = .17 und bei der Experimentalgruppe r = .05, wobei dieser Unterschied nicht signifikant ist. Für weitere Auswertungen unterteilten Mueller-Hanson et al. die beiden Gruppen anhand des Scores im LMI in je drei gleich grosse Untergruppen. Bei der Kontrollgruppe unterschieden sich die Korrelationen der hoch-scorenden von der tief-scorenden Untergruppen mit dem Score im Leistungstest nicht signifikant voneinander (r_{hoch} = .20, r_{tief} =. 26) bei der Experimentalgruppe jedoch schon (r_{hoch} = .07, r_{tief} =. 45). Abschliessend bestimmten sie die Anzahl der Studenten der Kontroll- und der Experimentalgruppe, welche auf Grund ihrer Ergebnisse im LMI in Abhängigkeit von der Selektionsquote die Selektionshürde geschafft hätten. Bei einer Selektionsquote von 20% (n = 91) stammen 62.6% der selektionierten Studienteilnehmern aus der Experimentalgruppe – welche 46% aller Studienteilnehmern stellen –, bei einer von 60% (n = 269) 54.0% und bei einer von 80% (n = 356) 49.0%. Diese Zahlen belegen, dass sich der Einfluss des Fakings bei einer select out-Strategie minimieren lässt.

Die anlässlich der Rekrutierung durchgeführte Kaderbeurteilung stellt – wie in Abbildung 1.1 ersichtlich – eine Negativ-Selektion dar (select out): Ziel ist nicht, einige wenige Top-Kandidaten ausfindig zu machen, sondern diejenigen Stellungspflichtigen zu bestimmen, welche das Potenzial für die Übernahme einer Kaderfunktion in der Armee *nicht* aufweisen. Dies, da sie auf Grund ihrer kogniti-

ven Leistungsfähigkeit und ihrer Persönlichkeitsmerkmale von einer Führungstätigkeit überfordert wären oder weil sie nicht in der Lage wären, die Verantwortung für andere Menschen zu übernehmen. Abbildung 1.1 zeigt auf, dass von den 65% als militärdiensttauglich beurteilten Stellungspflichtigen in einem ersten Schritt auf Grund der Testergebnisse ungefähr 40% eine Negativbeurteilung erhalten. Anlässlich des Zuteilungsgespräches vergibt der Rekrutierungsoffizier etwa zwei Dritteln der nicht negativ beurteilten Stellungspflichtigen eine positive Kaderempfehlung. Somit erhalten circa 40% der als militärdiensttauglich beurteilten Stellungspflichtigen eine positive Kaderempfehlung. Die Berufsoffiziere in den Rekrutenschulen schicken dann ungefähr 30% der Rekruten in eine Kaderschule, wobei es nicht zwingend ist, dass diese alle in der Rekrutierung eine positive Kaderempfehlung erhalten haben. Gemäss den Ergebnissen der Studie von Mueller-Hanson et al. muss also bei einer Selektionsquote von 60% auf Grund der Testergebnisse mit einer leichten Überrepräsentierung derjenigen Stellungspflichtigen gerechnet werden, welche sich beim Ausfüllen der Persönlichkeits-Fragebogen in einem stark positiven Licht dargestellt haben. Da bei diesem Selektionsschritt jedoch zusätzlich noch die Ergebnisse von zwei Intelligenztests in die Beurteilung einfliessen, in welchen sich der Stellungspflichtige nicht willentlich leistungsfähiger darstellen kann, wird der Effekt des Fakings nur noch schwach sein.

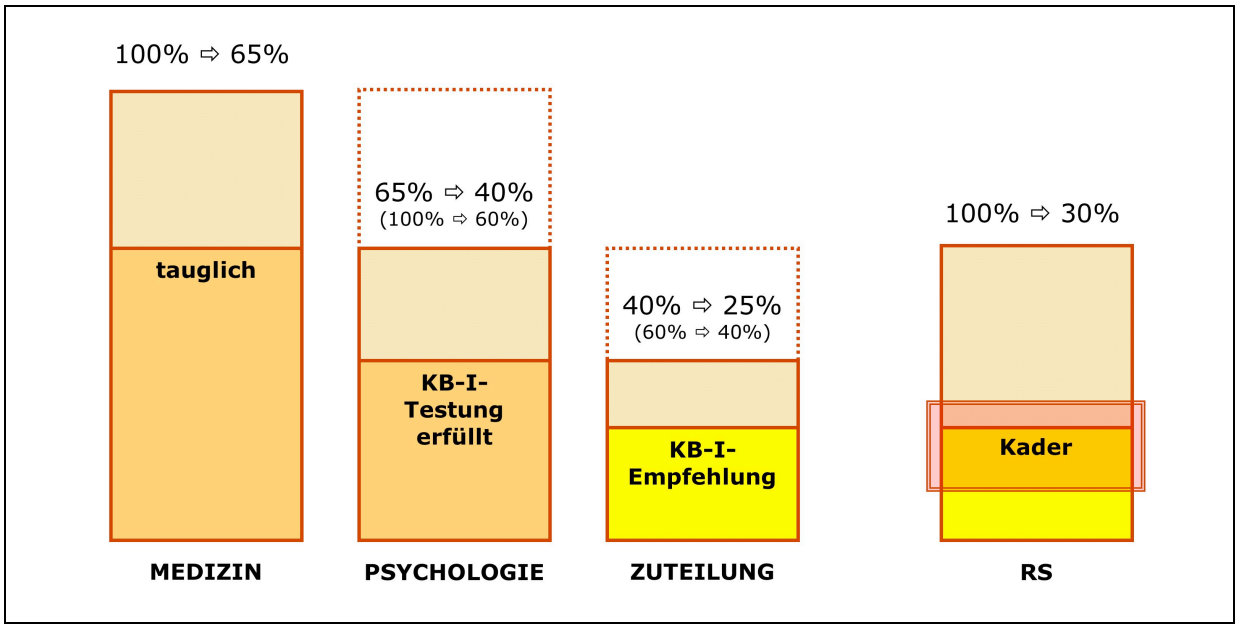


Abbildung 1.1 Ablauf und Zahlengerüst der Kaderbeurteilung Stufe I (Betschart, Boss & Jöri, 2009, S. 3)

An dieser Stelle weise ich noch kurz auf ein rekrutierungsspezifisches Phänomen hin, das *faking bad*, das bei Stellungspflichtigen auftritt, welche entweder keinen Militärdienst oder aber keinen Beförderungsdienst absolvieren möchten. So betragen die Korrelationen zwischen der Dienst- respektive der Führungsmotivation und den in der Rekrutierung eingesetzten Persönlichkeitsskalen zwischen r = .33 und .56 respektive r = .26 und .49 (Betschart, Boss & Jöri, 2009. Die Korrelationen zum Intelligenztest betragen jeweils r = -.04. N = 730 – 1'012. Siehe dazu auch Boss, König & Melchers, 2012.).

Während der Konzeptionsphase stellte sich die grundsätzliche Frage, ob für den Zweck der Rekrutierung bestehende Testverfahren – von kommerziellen Testanbietern oder fremden Armeen – eingekauft werden können oder ob neue Testverfahren zu entwickeln sind. In Tabelle 1.1 habe ich die Vor- und Nachteile des Einkaufs bestehender, kommerziell angebotener Testverfahren im Gegensatz zu einer Eigenentwicklung dargestellt.

Tabelle 1.1
Vor- und Nachteile des Einkauf bestehender Testverfahren und von Eigenentwicklungen (nach Stoll et al., 2000, S. 57)

	Einkauf bestehender Testverfahren	Eigenentwicklung
Vorteile	• schnell verfügbar • erprobte Anwendung • eigene, einfache Verfahren lassen sich relativ einfach ins System einfügen	• massgeschneiderte Testverfahren • die Tests lassen sich laufend an neue Bedürfnisse anpassen • alle drei Amtssprachen abgedeckt • in einem modernen, ansprechenden Design erstellbar • relativ tiefe Unterhaltskosten
Nachteile	• hohe, über die gesamte Einsatzzeit anfallende Kosten • Testverfahren z. T. nicht in allen drei Amtssprachen verfügbar • Abhängigkeit vom Testanbieter • bei Bekanntwerden der Testverfahren können sie geübt werden • evtl. Probleme bei der Verarbeitung grosser Datenmengen	• lange Konstruktionszeit • hohe einmalige Kosten • „Kinderkrankheiten"

Folgende drei Punkte waren bei der Entscheidungsfindung schliesslich ausschlaggebend:

- *Grosse Menge*: Pro Jahr absolvieren über 35'000 Stellungspflichtige die Rekrutierung, was beim Einsatz von Testverfahren kommerzieller An-

bieter hohe Kosten verursacht, so dass die Aufwände für Neuentwicklungen innert weniger Jahren amortisiert sein dürften.

- *Drei Sprachen*: Für die Rekrutierung müssen die Testverfahren in den drei Amtssprachen verfügbar sein. Dies stellt vor allem bei Persönlichkeits-Fragebogen ein Problem dar, da die kommerziellen Testanbieter über nur einige wenige Testverfahren verfügen, welche alle drei Sprachen abdecken.
- *Spezifische Fragestellungen*: Der Militärdienst stellt an Soldaten ganz spezifische Anforderungen, welche nur zum Teil mit denjenigen im Zivil- oder Berufsleben vergleichbar sind. Es ist davon auszugehen, dass die kommerziellen Testanbieter über keine Testverfahren verfügen, welche genau den gewünschten Anforderungen entsprechen, was jedoch für eine bestmögliche Aussagekraft von zentraler Bedeutung ist.

Die in diesem Kapitel dargestellten Ausführungen bezüglich der Kaderbeurteilung zusammenfassend lässt sich sagen, dass eine erste Einschätzung des Kaderpotenzials anlässlich der Rekrutierung grundsätzlich möglich ist. Dabei ist der gemeinsame, sich ergänzende Einsatz von Intelligenztests und Persönlichkeits-Fragebogen auf Grund der select out-Strategie zulässig und lässt aussagekräftige Ergebnisse erwarten. Da der Militärdienst ganz spezifische Anforderungen an Kaderangehörige stellt und im Hinblick auf die grosse Anzahl zu beurteilender Stellungspflichtiger und deren Heterogenität bezüglich Sprache und Schulbildung, drängt es sich auf, für diesen Zweck massgeschneiderte Testverfahren neu zu entwickelten.

Im nachfolgenden Kapitel gehe ich auf einige Besonderheiten ein, welchen ein Testverfahren genügen sollte, um im Rahmen der Rekrutierung eingesetzt werden zu können.

1.2 Anforderungen an die anlässlich der Rekrutierung eingesetzten Testverfahren

Schon bei der Planung der Rekrutierung A XXI zeichnete sich deutlich ab, dass in Zukunft die Stellungspflichtigen – also praktisch die Gesamtpopulation der Männer mit Schweizer Staatsbürgerschaft im Alter von 18 bis maximal 25 Jahren –

anlässlich ihrer Rekrutierung eine umfassende, in den drei Amtssprachen verfügbare, computergestützte Testbatterie absolvieren müssen. Dabei ergeben sich bedeutsame Unterschiede zu in der Privatwirtschaft durchgeführten Selektionsverfahren: Da die männlichen Stellungspflichtigen von ihrem Wohnortskanton ein verbindliches Aufgebot zur Rekrutierung erhalten, nehmen sie nicht freiwillig daran teil. Zudem hat bezüglich Kadereignung keinerlei Vorselektion stattgefunden und die Ärzte, Psychologen und Rekrutierungsoffiziere in den Rekrutierungszentren müssen Stellungspflichtige aus allen Bevölkerungsschichten und mit unterschiedlichsten Sozialisierungs- und Bildungshintergründen beurteilen. Aus diesem Grund formulierten Boss und Baumann (2003; siehe auch Boss, 2005) Kriterien, welche als Eckpfeiler für die Entwicklung der neuen Testverfahren dienten:

- Die Verfahren sollen von 19jährigen *gut akzeptiert* werden. Um dies zu erreichen sind die Inhalte der Fragebogen konsequent auf diese Altersgruppe abzustimmen und auf den Militärdienst auszurichten.
- Der Stellungspflichtige soll erkennen, zu welchem Zweck er die Fragebogen auszufüllen hat, indem er mündliche und schriftliche Informationen darüber erhält und die Testverfahren über eine *hohe Augenscheinvalidität* verfügen.
- Da die Stellungspflichtigen in den zwei bis drei Tagen insgesamt bis zu vier Stunden psychologische Testverfahren am Computer bearbeiten müssen, wird einerseits auf den Einsatz von *zeitökonomischen Verfahren* geachtet und andererseits auf eine *grosse Verfahrens- und Inhaltsvielfalt*, um die Ermüdung in Grenzen zu halten und Langeweile zu verhindern.
- Die Testverfahren müssen in einer *klaren, einfachen Sprache* verfasst sein, damit sie von den meisten der Stellungspflichtigen problemlos verstanden werden und um die Übersetzung in die anderen Amtssprachen zu vereinfachen.

Nicht aufgeführt sind hier die Anforderungen, welche grundsätzlich und unabhängig von deren Einsatzzweck an Testverfahren gestellt werden, wie die Hauptgütekriterien Objektivität, Reliabilität und Validität und die Nebengütekriterien wie zum Beispiel Normierung, Vergleichbarkeit, Nützlichkeit oder Fairness (z. B. Kubinger & Proyer, 2005; Lienert & Raatz, 1998). Die oben aufgelisteten Punkte betreffen – im Gegensatz zu den „technischen" Anforderungen an Testverfahren – eher die sozialen Aspekte einer Testdurchführung (Schuler & Stehle, 1983) oder Kriterien einer bewerberzentrierten Personalselektion (Boss, 2005). Gilliland hat 1993 auf der Grundlage der Ergebnisse aus der Gerechtigkeitsfor-

schung ein Modell der Bewerberreaktionen auf Personalauswahlverfahren erstellt, welches folgende zehn Regeln umfasst: Tätigkeitsbezug, Möglichkeit zur Selbstdarstellung, Möglichkeit zur Wiedererwägung, Vergleichbarkeit der Durchführung, Ergebnisrückmeldung, Information zum Auswahlverfahren, Aufrichtigkeit, respektvolle Behandlung, Zweiweg-Kommunikation, Angemessenheit der Fragen. Ich stelle dieses Modell ausführlich in Kapitel 5.4 dar. Da der Einsatz von Testverfahren im Rahmen der staatlich angeordneten Rekrutierung, an welcher praktisch alle jungen Schweizer Bürger teilnehmen müssen, eine aussergewöhnliche Situation darstellt, habe ich – ausgehend von oben aufgeführter Liste – weitere drei Aspekte einer bewerberzentrierten Personalselektion formuliert: Verhältnismässigkeit, Adressatspezifität und Ansprechcharakter.

Verhältnismässigkeit

Die Teilnahme an einem Personalauswahlprozess ganz allgemein und an der Rekrutierung im Speziellen stellt für die meisten Bewerber respektive Stellungspflichtigen eine belastende Situation dar. Aus diesem Grund ist anhand eines differenzierten Anforderungsprofils genau zu definieren, welche Informationen der Rekrutierungsarzt, der Psychologe und der Rekrutierungsoffizier für deren zu treffende Entscheidungen benötigen und wie diese möglichst zeitökonomisch und wenig Stress auslösend erhoben werden können. Die Antwort auf die Frage nach dem optimalen Verhältnis zwischen Aufwand und Ertrag lässt sich dabei nicht nur anhand organisatorischer Gesichtspunkte finden, sondern es muss auch das Erleben der Situation durch die Stellungspflichtigen einbezogen werden. Die Verhältnismässigkeit lässt sich auch wahren, indem man den Selektionsprozess sequenziell gestaltet, so dass nur noch eine Gruppe von vorselektionierten Stellungspflichtigen anstrengende oder zeitintensive Testungen absolviert (*sequenzielles Selektionsverfahren*) oder dass man massgeschneiderte (sog. *tailored tests*) oder adaptive Testverfahren einsetzt. Auch in der DIN 33430 zu den Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen (2002, S. 13) ist dieser Aspekt enthalten: „Die Kandidaten sollen zeitlich, psychisch und körperlich nicht mehr beansprucht werden als es für den Untersuchungszweck erforderlich ist."

Adressatspezifität

Beim Einsatz von Testverfahren im Rahmen der Rekrutierung ist darauf zu achten, dass die Inhalte, die Form und das Anspruchsniveau auf die Stellungspflichtigen abgestimmt sind. Aus diesem Grund ist es auch nur bedingt möglich, Verfahren, welche für die Personalselektion in der Wirtschaft zum Einsatz gelangen,

ohne entsprechende Modifikationen einzusetzen. Ein Dilemma ergibt sich vor allem bei Leistungstests, bei deren Bearbeitung sich die Stellungspflichtigen weder unter- noch überfordert fühlen sollten. Da an der Rekrutierung jedoch praktisch die gesamte männliche Jugend teilnimmt, sieht man sich auch mit dem gesamten Leistungsspektrum konfrontiert. Eine Lösung für dieses Dilemma könnten auch hier adaptive Testverfahren sein, bei welchen das Testsystem jedem Stellungspflichtigen diejenigen Items zur Bearbeitung vorlegt, welche dieser mit einer Chance von 50% richtig löst.

Ansprechcharakter

Da die Stellungspflichtigen in den Rekrutierungszentren mehrere Fragebögen zu verschiedenen Themenbereichen bearbeiten müssen, ist darauf zu achten, dass sich die einzelnen Verfahren bezüglich Aufgabencharakter und Layout unterscheiden, so dass bei der Bearbeitung weder Langeweile noch Überdruss entsteht. Bei computergestützten Testverfahren darf zudem eine ansprechende Präsentationsweise erwartet werden. Es ist anzunehmen, dass dies sowohl Auswirkungen auf die Motivation der Stellungspflichtigen, den Test zu bearbeiten, als auch auf die Einschätzung der Zuverlässigkeit des Verfahrens hat.

Diese drei Aspekte decken sich teilweise mit dem Kriterium Tätigkeitsbezug (Ausrichtung auf den Militärdienst) aus dem Modell von Gilliland (1993) und dem Kriterium der Transparenz (Augenscheinvalidität) aus dem Konzept der sozialen Validität von Schuler und Stehle (1983). Bei der Entwicklung des Leadership-Fragebogens entschied ich mich trotzdem, das Kriterium Tätigkeitsbezug nur teilweise umzusetzen: Erstens sollten die im Leadership-Fragebogen geschilderten Situationen möglichst aus dem Erfahrungsschatz der Stellungspflichtigen stammen. Bei der Verwendung militärischer Situationen wäre dies nicht der Fall gewesen, das heisst die Stellungspflichtigen hätten sich nicht ähnliche Situationen, die sie selbst schon erlebt haben, in Erinnerung rufen können, sondern hätten rein hypothetisch antworten müssen, wie sie sich in der geschilderten Situation verhalten würden. Zudem wollte ich verhindern, dass einzelne der Stellungspflichtigen, welche unter keinen Umständen Militärdienst leisten wollen, mit Reaktanz auf die militärischen Situationen und die Testung insgesamt reagieren. Wie zentral der Aspekt der Akzeptanz der in den Rekrutierungszentren eingesetzten Testverfahren ist, zeigte sich kurz nach dem Start der neuen Rekrutierung im Sommer 2003, als die Boulevard-Presse auf Grund von Motionen und Interpellationen im Nationalrat die zum Teil sehr persönlichen und intimen Fragen des Medizin-psychologischen (psychiatrischen) Fragebogens an den Pranger stellten:

- Motion vom 19.6.2003 von Nationalrätin Franziska Teuscher: *Aushebung. Keine Schnüffelei.* Der Bundesrat wird beauftragt, 1. den Fragebogen für die militärische Aushebung der Rekruten überarbeiten zu lassen und alle Fragen, welche in die Privat- und Persönlichkeitssphäre eingreifen, aus dem Test zu streichen. ...
- Interpellation vom 20.6.2003 von Nationalrat Alexander Baumann: *Fragebogen bei der militärischen Aushebung* ... 3. Ist der Bundesrat der Ansicht, dass Fragen zum Sexualleben zur Beurteilung der Diensttauglichkeit etwas beitragen? ...

Die Boulevard-Zeitung Blick schlachtete dieses Thema anschliessend aus:

- „Schnüffel-Schweinerei beim Militär. 400 Skandalfragen bei der Aushebung – die geheime Liste." (Blick, 02.07.2003)
- „Ungesetzlich! Schluss mit der Schnüffel-Schweinerei beim Militär." (Blick, 03.07.2003)
- „VBS-Skandalfragen im Internet aufgetaucht. Schnüffeltest nichts mehr wert?" (Blick, 04.07.2003)
- „Schnüffel-Schweinerei beim Militär. VBS kippt Sexfragen." (Blick, 05.07.2003)

Den krönenden Abschluss fand diese Affäre Ende 2003, als die Organisation „Big Brother Award Schweiz" dem damaligen Bundesrat des Departements für Verteidigung, Bevölkerungsschutz und Sport eine Auszeichnung als eifrigster Schnüffler der Schweiz verlieh. Die ganze Aufregung hätte sich leicht vermeiden lassen, wenn alle Testverfahren anhand der Kriterien einer bewerberzentrierten Selektion entwickelt worden wären. So habe ich bereits im Konzept zu den psychischen Aspekten der Rekrutierung A XXI (Stoll, Boss & de With, 2000) bei meinen Ausführungen zur Erfassung der psychischen Ressourcen im Rahmen der Diensttauglichkeitsabklärung darauf hingewiesen, dass sich viele der klinisch orientierten Fragebogen nicht für das Screening der Stellungspflichtigen eignen:

> Die einzelnen Fragen sollen klinisch orientiert, aber auf den Militärdienst zugeschnitten sein. Das heisst, dass nur Fragen und Aussagen aufgenommen werden, welche in einem eindeutigen Bezug zum Militärdienst stehen und die Privatsphäre der Stellungspflichtigen so weit wie möglich wahren. So ist es nicht zulässig, Aussagen im Stil wie
>
> „Mein Sexualleben ist zufriedenstellend."
>
> „Meine Seele verlässt manchmal meinen Körper."
>
> „Ich glaube, ich bin ein verdammter Mensch."

beantworten zu lassen. Da solche Aussagen in gängigen, klinischen Persönlichkeitsinventaren jedoch vorkommen, ist dies zugleich der Hauptgrund, weshalb ein neuer Fragebogen für die Rekrutierung erstellt werden muss. (Stoll et al., 2000, S. 23)

Abschliessend zu diesem Kapitel stelle ich in Tabelle 1.2 die im Jahre 2008 in der Rekrutierung eingesetzten psychologischen Testverfahren dar. Für 85% der Stellungspflichtigen beträgt die reine Testzeit weniger als 2.5 Stunden, leseschwache Stellungspflichtige benötigen bis zu vier Stunden, um alle Tests zu absolvieren.

Tabelle 1.2
Anlässlich der Rekrutierung eingesetzte Testverfahren (Stand 2008)

Testverfahren	Einsatzgebiet			Anzahl Items	Zeitbedarf (PR 85)
	DT	Fkt	KB I		
Fragebogen zu psycho-sozialen Belastungen	X			147	22.0
Medizin-psychologischer Fragebogen	X			252	29.9
Intelligenztest 95	X	X	X	2 x 30	15.0
Textverständnistest	X		X	2 x 8	14.7
Sportliche Leistungsprüfung (Fragebogen)		X		21	3.0
Interessen-Inventar		X		105	7.0
Persönlichkeits-Fragebogen		X	X	80	10.1
Leadership-Fragebogen		X	X	30	18.8
Merkfähigkeits-Test		X	X	2 x 18	26.5
Führungsmotivations-Fragebogen			X	11	1.4
Eignungsprüfung für Motorfahrer		*X*			*90 - 210*

Anmerkung. *N* = 49'000 – 57'000. Datensatz der Jahre 2007 / 08. DT = Diensttauglichkeitsabklärung, Fkt = Funktionszuteilung, KB I = Kaderbeurteilung Stufe I. Zeitbedarf PR 85 = Zeitdauer in Minuten, in welcher 85% der Stellungspflichtigen den Test absolviert haben (PR = Prozentrang).

1.3 Grobkonzept für die Entwicklung des Leadership-Fragebogens für die Kaderselektion anlässlich der Rekrutierung

Die Ausgangslage für die Entwicklung eines Testverfahrens für den Einsatz in der Kaderbeurteilung anlässlich der Rekrutierung lässt sich auf Grund oben dargestellter Ausführungen wie folgt charakterisieren:

- Die Projektleitung „Rekrutierung A XXI" fordert eine erste Erfassung des Kaderpotenzials anlässlich der Rekrutierung.
- Die spezifischen Rahmenbedingungen der Rekrutierung und die geforderte Dreisprachigkeit legen die Neuentwicklung eines Testverfahrens nahe.
- Um eine möglichst hohe Akzeptanz zu erzielen, ist die neu zu entwickelnde Testung der Population der Stellungspflichtigen anzupassen.
- Da die Stellungspflichtigen mehrere Testverfahren absolvieren müssen, sollten sich diese bezüglich Inhalt und Erscheinungsbild deutlich unterscheiden, um so einem Testüberdruss vorzubeugen.

Es stellt sich nun die Frage, welche Testverfahren für den Einsatz in der Kaderbeurteilung der Rekrutierung geeignet wären. Dazu muss jedoch zuerst festgelegt werden, welche Kompetenzen und Persönlichkeitseigenschaften dabei zu erheben und hinsichtlich des Kaderpotenzials zu beurteilen sind. In Kapitel 6.1 stelle ich ausführlich die Entwicklung des Anforderungsprofils für untere Kader der Schweizer Armee dar. Um die Ausgangslage für die Konstruktion des Leadership-Fragebogens vollständig darzulegen, führe ich an dieser Stelle das für das Konzept zu den psychologischen Aspekten der neuen Rekrutierung (Stoll et al., 2000) erstellte provisorische Anforderungsprofil auf. Anhand des Studiums empirischer Arbeiten zu den Anforderungen an militärische Führungskräfte in der Schweizer Armee (Annen, 2000; Hoenle, 1996; Stadelmann, 1998) ergaben sich folgende Dimensionen, welche für die anlässlich der Rekrutierung stattfindende Kaderbeurteilung bedeutsam sind:

- Intelligenz (Konzentration, Ausdauer, Belastbarkeit)
- Persönlichkeitsdimensionen (interne Kontrollüberzeugung, Leistungsmotivation, Durchsetzungsvermögen, Durchhaltevermögen, Gewissenhaftigkeit, Integrität, emotionale Stabilität)
- Soziale Kompetenzen (Führungsmotivation, Teamfähigkeit, Konfliktfähigkeit, Frustrationstoleranz, Dominanzstreben)

Der neu zu entwickelnde Leadership-Fragebogen – im Konzept noch als Test zur Erfassung der sozialen Kompetenz I bezeichnet – deckt den Bereich der sozialen Kompetenzen ab und soll Teamfähigkeit, Konfliktfähigkeit, Frustrationstoleranz und Dominanzstreben messen. (Für die Erfassung der Führungsmotivation habe ich in Zusammenarbeit mit Studierenden einen Kurzfragebogen entwickelt.) Ziel ist es somit nicht, die Persönlichkeit umfassend abzubilden, wie dies zum Beispiel mit einem Fragebogen zu den Big Five möglich wäre, sondern einige zentrale Persönlichkeitsmerkmale für unteres Kader in der Schweizer

Armee zu operationalisieren. Da soziale Kompetenzen definitionsgemäss in Situationen im Umgang mit Menschen gefordert sind, habe ich mich für einen situativen Test für deren Erfassung entschieden. Dabei werden dem Stellungspflichtigen Situationen mit sozialen Interaktionen mit jeweils zwei oder mehreren vorgegebenen Verhaltensalternativen geschildert. Der Stellungspflichtige hat die Aufgabe, diejenige Verhaltensalternative auszuwählen, welche er in der jeweiligen Situation am ehesten zeigen würde. Ich habe mich dabei an eine Verfahrenskategorie aus dem Testsystem "Professional Assessment by Computer for Training and Selection - profacts" von Etzel und Hornke angelehnt. Die folgende Abbildung 1.2 zeigt ein Item aus dem Instrument "Sales and Communication", welches – mit einer passenden Fotografie illustriert – eine Situation zur Erfassung der sozialen Flexibilität beschreibt. Wie bei herkömmlichen Persönlichkeits-Fragebogen handelt es sich bei diesem – als Situational Judgment Test bezeichneten (z. B. Lievens, Peeters & Schollaert, 2008; McDaniel & Whetzel, 2007; Weekley & Ployhart, 2006) – Testtyp auch um eine Selbstbeschreibung, mit dem Unterschied, dass die Situation, in welcher das Verhalten beschrieben werden soll, ausführlich dargestellt wird. Somit sind im Rahmen der Testentwicklung zu den vier zu erfassenden Dimensionen passende Alltagssituationen aus dem Leben 20jähriger zu formulieren.

Abbildung 1.2 Beispiel eines situativ verankerten Items aus dem profacts-Testsystem (Etzel & Küppers, 2002, S. 145).

Bei der Entwicklung eines neuen Testverfahrens stellt sich die grundsätzliche Frage, wie man zum Itempool gelangt. Bei der wohl am häufigsten eingesetzten Vorgehensweise zieht sich der Testentwickler oder das Entwicklungsteam in ein stilles Kämmerchen zurück und entwickelt auf der Grundlage der Definitionen der Testdimensionen den Itempool. Hilfreich sind dabei sicherlich eine langjährige Erfahrung bei der Itementwicklung und ein gewisses Flair für diese Aufgabe. So schreibt zum Beispiel Osterlind (1998, S. 1–2) vom „skilled test developer", welcher sich einen „sixth sense" für die Testitemkonstruktion erarbeiten muss. Und auch Nunnally und Bernstein (1994) sehen in der Entwicklung von Testitems eine Kunst:

> A good plan provides an intention to construct a good test, but unless items are skillfully written, the plan never materializes. Although there are some rules for writing good items ..., writing test items is an art that few people master. (S. 297)

Nicht alle Testentwickler teilen jedoch diese Ansicht: So soll Hornke schon in den 70er Jahren zu Cronbach gesagt haben, dass „some day [items] may be computer generated. ... Items should be based on theory and constructs derived from this theory" (Hornke, 2002, S. 159). Er selbst hat dies dann auch so umgesetzt, indem er seine Tests regel- oder theoriegeleitet entwickelt hat (Hornke, 2002; Hornke, Küppers & Etzel, 2000; Hornke, Rettig & Hutwelker, 1988), eine Vorgehensweise, die im deutschen Sprachraum auf gute Resonanz gestossen ist (z. B. Altstötter-Gleich, 1998; Etzel, 1999; Gittler & Arendasy, 2003). Grundlage dafür ist die Entwicklung eines Konstruktionsrationals, welches anhand konkreter Regeln beschreibt, wie bei der Generierung der Items vorzugehen ist. Eine konsequente Weiterführung des Gedankens der regelgeleiteten Testentwicklung ist die automatisierte Itemgenerierung, bei welcher der Computer im Verlauf der Testung anhand eines Baukastens an Itembestandteilen neue Items konstruiert, welche exakt den für den Testbearbeiter angepassten Schwierigkeitsgrad haben (Arendasy, Sommer, Gittler & Hergovich, 2006; Bejar, Lawless, Morley, Wagner, Bennett & Revuelta, 2003; siehe auch Wainer, 2002). Dieses Verfahren lässt sich jedoch nur auf Leistungstests anwenden und hat sich wohl auf Grund der Komplexität bei der Umsetzung bis heute nicht durchsetzen können.

Auf Grund des Auftrages der Projektleitung „Rekrutierung A XXI" und der oben aufgeführten Überlegungen zu den Anforderungen an ein in der Rekrutierung einzusetzendes Testverfahren, lässt sich ein Grobkonzept für den zu erstellenden Leadership-Fragebogen ableiten. Ich habe dieses in drei Aussagen verdichtet, welche auch gleichzeitig die Ziele darstellen, die ich erreichen möchte:

1. Der als Situational Judgment Test konzipierte Leadership-Fragebogen erfasst a priori definierte Persönlichkeitsdimensionen.
2. Die Fragebogenentwicklung basiert auf einem Konstruktionsrational.
3. Die Stellungspflichtigen akzeptieren den Leadership-Fragebogen gut.

Für diese Arbeit stehen also das Vorgehen bei der Testentwicklung und die Akzeptanz des Leadership-Fragebogens im Zentrum des Forschungsinteresses. Ich untersuche, ob sich mit dem gewählten Itemlayout und einer rationalen Testkonstruktion ein gut akzeptierter Persönlichkeits-Fragebogen entwickeln lässt. Damit habe ich deutlich andere Aspekte in den Vordergrund gerückt, als dies sonst bei der Entwicklung psychologischer Testverfahren üblich ist. Unbestritten bleibt jedoch, dass die prädiktive Validität bei einem Testverfahren, welches in der Personalselektion eingesetzt wird, das wichtigste Gütekriterium darstellt und diese auch zu überprüfen ist.

1.4 Ausblick auf die vorliegende Arbeit

Das zu den drei Aussagen verdichtete Grobkonzept des zu entwickelnden Leadership-Fragebogen findet seinen Niederschlag auch in den Themen, welche ich im Theorieteil zu dieser Arbeit behandle: Situational Judgment Tests, Konstruktionsrationale (Act Frequency Approach und Wertequadrat) und Akzeptanz von Testverfahren im Rahmen der Personalselektion:

In Kapitel 2 gehe ich auf die Grundlagen von Situational Judgment Tests ein und beschreibe anschliessend ausführlich die für die Entwicklung eines solchen Testverfahrens notwendigen Schritte. Die Darstellung der Zusammenhänge von Situational Judgment Tests mit Arbeitsleistung, Intelligenz und Persönlichkeit schliesst dieses Kapitel ab.

Die beiden nachfolgenden Kapitel 3 und 4 behandeln die theoretischen Fundamente der Konstruktion des Leadership-Fragebogens, den Act Frequency Approach und das Wertequadrat. Bei beiden Themen gehe ich zuerst auf die Grundlagen und den historischen Hintergrund ein und stelle anschliessend anhand konkreter Beispiele dar, wie diese beiden Ansätze für die Konstruktion von Persönlichkeits-Fragebogen eingesetzt werden können.

Den Abschluss des Theorieteils dieser Arbeit bildet ein Kapitel über die Akzeptanz von Testverfahren. Ich beginne mit einer Darstellung der Krise in der psychologischen Diagnostik, welche im deutschen Sprachraum den Ausgangs-

punkt für eine wissenschaftliche Auseinandersetzung mit den Reaktionen der Testbearbeiter auf die Testung bildete und zum Konzept der sozialen Validität von Schuler und Stehle (1983) führte. Ausführlich stelle ich das Modell der Bewerberreaktionen auf Personalauswahlverfahren von Gilliland (1993) dar, welches umfangreiche Forschungsaktivitäten auslöste. Anschliessend gehe ich auf verschiedene Skalen zur Erfassung der Akzeptanz von Testverfahren ein und schliesse das Kapitel mit einer Übersicht zur Akzeptanz verschiedener im Bereich der Personalselektion eingesetzter Testverfahren ab.

Den Ergebnisteil habe ich auf zwei Kapitel aufgeteilt: In Kapitel 6 beschreibe ich das Vorgehen bei der Konstruktion des Leadership-Fragebogens auf der Basis der für die drei Dimensionen Durchsetzungsfähigkeit, Kontaktfähigkeit und Verantwortungsbewusstsein entwickelten Wertequadrate und mit dem Einsatz des Act Frequency Approachs. Zu Beginn dieses Kapitels stelle ich die von mir durchgeführten Anforderungsanalysen für unteres Milizkader der Schweizer Armee dar und ordne die Dimensionen des Leadership-Fragebogens in das Anforderungsprofil für Gruppenführer ein.

In Kapitel 7 beschreibe ich die Studien zur Überprüfung der Faktorenstruktur einer gekürzten Version des Leadership-Fragebogens. Weiter untersuche ich die Auswirkungen unterschiedlicher Scoring-Arten auf die Reliabilität der Leadership-Skalen, stelle die Studie zum Bekanntheitsgrad der Items vor und berichte von den Ergebnissen der Überprüfung der Akzeptanz des Leadership-Fragebogens.

Den Abschluss dieser Arbeit bildet die Diskussion der Testentwicklung und der Erkenntnisse aus den Überprüfungsstudien.

An dieser Testentwicklung haben – wie im Vorwort schon erwähnt– viele Studentinnen und Studenten und Projektangestellte mitgearbeitet. Wenn im folgenden Text von „wir" die Rede ist, will ich damit darauf hinweisen, dass bei diesem Entwicklungsschritt ein unter meiner Leitung stehendes Team an der Arbeit war.

1.5 Literaturverzeichnis

Alliger, G. M., Lilienfield, S. O., & Mitchell, K. E. (1995). The susceptibility of overt and covert integrity tests to coaching and faking. *Psychological Science, 7*, 32–39.

Altstötter-Gleich, C. (1998). Theoriegeleitete Itemkonstruktion und –auswahl mittels der Repertory-Grid-Technik. *Zeitschrift für Differenzielle und Diagnostische Psychologie, 19*, 149–163.

Annen, H. (2000). *Förderwirksame Beurteilung. Aktionsforschung in der* Schweizer Armee. Frauenfeld: Huber.

Arendasy, M., Sommer, M., Gittler, G., & Hergovich, A. (2006). Automatic generation of quantitative reasoning items. A pilot study. *Journal of Individual Differences, 27*, 2–14.

Asendorpf, J. B. (1999). *Psychologie der Persönlichkeit* (2. Auflg.). Berlin: Springer.

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1–26.

Barrick, M. R., & Mount, M. K. (2005). Yes, personality matters: Moving on to more important matters. *Human Performance, 18*, 359–372.

Beaty, J. C., Cleveland, J. N., & Murphy, K. R. (2001). The relation between personality and contextual performance in "strong" versus "weak" situations. *Human Performance, 14*, 125–148.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *The Journal of Technology, Learning and Assessment, 2*, 3–29.

Bertua, C., Anderson, N., & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology, 78*, 387–409.

Betschart, I., Boss, P. & Jöri, E. (2009). *Bestandesaufnahme der Kaderbeurteilung Stufe I* (Unveröffentlichter Bericht). Zürich: Universität Zürich, Psychologisches Institut, Fachrichtung Sozial- und Wirtschaftspsychologie.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on per-

sonality measures. *International Journal of Selection and Assessment, 14*, 317–335.

Boss, P. (2005). Assessment in der Arbeitswelt – Kriterien für eine bewerberzentrierte Personalauswahl. In M. Rehbinder (Hrsg.), *Psychologische Aspekte im Recht der Personalführung* (S. 21–45). Bern: Stämpfli.

Boss, P. & Baumann, R. (2003). Psychologische Testverfahren beim Rekrutierungsprozess der Armee. *HR-Today, 7–8*, 22–23.

Boss, P., König, C. J., & Melchers, K. G. (2012). *Faking good and faking bad among army conscripts.* Manuscript submitted for publication.

Boss, P., Vetter, S., Frey, F. & Lupi, G. A. (2003). Rekrutierung XXI. 2. Teil: Die medizinisch-psychologischen und die psychologischen Testserien und Untersuchungen an der Rekrutierung XXI. *Schweizerische Ärztezeitung, 84*, 623–627.

Büttiker, W. & Stoller, S. (1989). *Erfahrungen junger Instruktoren mit Kaderauswahl und Dienstbetrieb.* Unveröffentlichte Seminararbeit an der Militärischen Führungsschule.

Conley, J. J. (1984). Longitudinal consistency of adult personality: Self-reported psychological characteristics across 45 years. *Journal of Personality and Social Psychology, 47*, 1325–1333.

DIN (2002). *DIN 33430: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen*. Berlin: Beuth.

Dunlap, W. P., & Cornwell, J. M. (1994). Factor analysis of ipsative measures. *Multivariate Behavioral Research, 29*, 115–126.

Etzel, S. (1999). *Multimediale, computergestützte diagnostische Verfahren: Neue Perspektiven für die Managementdiagnostik.* Aachen: Shaker.

Etzel, S. & Küppers, A. (2002). Mit Methodenvielfalt zum Ziel – Computergestützte und klassische Assessmenttechniken. In R. Bäcker & S. Etzel (Hrsg.). *Einzel-Assessment. Neue Verfahren zur Auswahl und Entwicklung von Führungskräften* (S. 135–171). Düsseldorf: Symposion.

Frey, F., Huber, R. & Lupi, G. A. (2003). Rekrutierung XXI. Der medizinische Teil der Rekrutierung XXI. Übersicht und aktueller Stand dieses Projektes. Einbezug der Zivilärzte. *Schweizerische Ärztezeitung, 84*, 341–345.

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *The Academy of Management Review, 18*, 694–734.

Gittler, G. & Arendasy, M. (2003). Endlosschleifen: Psychometrische Grundlagen des Aufgabentyps E^p. *Diagnostica, 49*, 164–175.

Griffith, R. L., Chmielowski, T. S., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 3*, 341–355.

Grubitzsch, S. & Rexilius, G. (Hrsg.). (1978). *Testtheorie – Testpraxis. Voraussetzungen, Verfahren, Formen und Anwendungsmöglichkeiten psychologischer Tests im kritischen Überblick*. Reinbek: Rowohlt.

Guion, R. M., & Gottier, R. F. (1965). Validity of personality measures in personnel selection. *Personnel Psychology, 18*, 135–164.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*, 167–184.

Hoenle, S. (1996). *Führungskultur in der Schweizer Armee.* Frauenfeld: Huber.

Hogan, R. (2005). In defense of personality measurement: New wine for old whiners. *Human Performance, 18*, 331–341.

Hornke, L. F. (2002). Item-generation models for higher order cognitive functions. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 159–178). Mahwah, NJ: Erlbaum.

Hornke, L. F., Küppers, A. & Etzel, S. (2000). Konstruktion und Evaluation eines adaptiven Matrizentests. *Diagnostica, 46*, 182–188.

Hornke, L. F., Rettig, K. & Hutwelker, R. (1988). Theoriegeleitete Konstruktion eines Tests zur Messung des räumlichen Vorstellungsvermögens. In Bundesministerium für Verteidigung (Hrsg.), *Untersuchungen des Psychologischen Dienstes der Bundeswehr, 23*, 145–222.

Hough, L. M., & Furnham, A. (2003). Use of personality variables in work settings. In I. B. Weiner (Ed.) & W. Borman, D. Illgen, & R. Klimoski (Vol. Eds.), *Comprehensive handbook of psychology Vol. 12. Industrial and organizational psychology* (pp. 131–169). New York, NY: Wiley.

Hough, L. M., & Oswald, F. L. (2005). They're right, well … mostly right: Research evidence and an agenda to rescue personality testing from 1960s insights. *Human Performance, 18*, 373–387.

Hurtz, G. M., & Donovan, J. J. (2000). Personality and job performance: The Big Five revisited. *Journal of Applied Psychology, 85*, 869–879.

Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology, 61*, 153–162.

Judge, T. A., Bono, J. E., Ilies, R., & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology, 87*, 765–780.

Kubinger, K. D. & Proyer, R. (2005). Gütekriterien. In K. Westhoff, L. J. Hellfritsch, L. F. Hornke, K. D. Kubinger, F. Lang, H. Moosbrugger, A. Püschel & G. Reimann (Hrsg.), *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (2., überarb. Auflg.) (S. 191–199). Lengerich: Pabst Science Publishers.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Psychologie Verlags Union.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review, 37, 426–441*.

Marcus, B. (2003). Persönlichkeitstests in der Personalauswahl: Sind "sozial erwünschte" Antworten wirklich nicht wünschenswert? *Zeitschrift für Psychologie, 211*, 138–148.

McDaniel, M. A., & Whetzel, D. L. (2007). Situational judgment tests. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 235–257). Hillsdale, NJ: Erlbaum.

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and Organizational Psychology, 77*, 531–552.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007a). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683–729.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007b). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology, 60*, 1029–1049.

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C., III. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology, 88*, 348–355.

Murphy, K. R., & Dzieweczynski, J. L. (2005). Why don't measure of broad

dimensions of personality perform better as predictors of job performance? *Human Performance, 18*, 343–357.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory.* New York, NY: McGraw-Hill.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*, 995–1027.

Ones, D. S., & Viswesvaran, C. (2001). Personality at work: Criterion-focused occupational personality scales used in personnel selection. In B. W. Roberts & R. Hogan (Eds.), *Personality psychology in the workplace* (pp. 63–92). Washington, DC: American Psychological Association.

Ones, D. S., Viswesvaran, C., & Dilchert, S. (2005). Personality at work: Raising awareness and correction misconceptions. *Human Performance, 18*, 389–404.

Osterlind, S. J. (1998). *Constructing test items: Multiple-Choice, constructed-response, performance, and other formats* (2nd ed.). Boston, MA: Kluwer Academic Publishers.

Pulver, U., Lang, A. & Schmid, F. W. (Hrsg.). (1978). *Ist Psychodiagnostik verantwortbar? Wissenschaftler und Praktiker diskutieren Anspruch, Möglichkeiten und Grenzen psychologischer Erfassungsmittel.* Bern: Huber.

Rossé, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion of preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83*, 634–644.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review, 16*, 155–180.

Salgado, J. F. (1997). The five factor model of personality and job performance in the European community. *Journal of Applied Psychology, 82*, 30–43.

Schmidt, F. L., & Hunter, J. E. (1998a). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schmidt, F. L. & Hunter, J. E. (1998b). Messbare Personenmerkmale: Stabilität, Variabilität und Validität zur Vorhersage zukünftiger Berufsleistung und berufsbezogenen Lernens. In M. Kleinmann & B. Strauss (Hrsg.), *Potentialfeststellung und Personalentwicklung* (S. 15–43). Göttingen: Verlag für Angewandte Psychologie.

Schuler, H. & Stehle, W. (1983). Neuere Entwicklungen des Assessment-Center-Ansatzes – beurteilt unter dem Aspekt der sozialen Validität. *Psychologie und Praxis. Zeitschrift für Arbeits- und Organisationspsychologie, 27*, 33–44.

Schweizerische Armee (2000). *Konzeptionsstudie 6 A. Rekrutierung A XXI.* Bern: Selbstverlag.

Schweizerische Gesellschaft für Psychologie (1975). Jahresversammlung 1975. Symposium Krise der Diagnostik. *Schweizerische Zeitschrift für Psychologie, 34*, 205–249.

Schweizerische Gesellschaft für Psychologie (1976). Krise der Diagnostik. Fortsetzung der Diskussion. *Schweizerische Zeitschrift für Psychologie, 35*, 49–61.

Stadelmann, J. (1998). *Führung unter Belastung. Ausgewählte Aspekte der Militärpsychologie.* Frauenfeld: Huber.

Stoll, F., Boss, P. & de With, E. (2000). *Konzept Rekrutierung Armee XXI. Psychologische Aspekte.* Zürich: Universität Zürich, Abteilung Angewandte Psychologie.

Tenopyr, M. L. (1988). Artifactual reliability of forced-choice scales. *Journal of Applied Psychology, 73*, 749–751.

Tett, R. P., Anderson, M. G., Ho, C. L., Yang, T. S., Huang, L., & Hanvongse, A. (2006). Seven nested questions about faking on personality tests: An overview and interactionist model of item-level response distortion. In R. L. Griffith & M. H. Peterson (Eds.), *A closer examination of applicant faking behavior* (pp. 43–83). Greenwich, CT: Information Age Publishing.

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*, 500–517.

Tett, R. P., & Christiansen, N. D. (2007). Personality tests at the crossroads: A response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt (2007). *Personnel Psychology, 60*, 967–993.

Wainer, H. (2002). On the automatic generation of test items: Some whens, whys, and hows. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 287–305). Mahwah, NJ: Erlbaum.

Weekley, J. A., & Ployhart, R. E. (Eds.). (2006). *Situational judgment tests: Theory, measurement, and application*. Mahwah, NJ: Erlbaum.

2. Situational Judgment Tests

2.1 Charakterisierung der Situational Judgment Tests

> Situational judgment tests ... are a popular selection method that present applicants with work related situations and ask them to indicate how they would respond to each situation. (Ployhart & Ehrhart, 2003, S. 1)

Situational Judgment Tests (SJTs) bilden eine Klasse psychodiagnostischer Verfahren respektive eine spezifische Messmethode, die sich dadurch charakterisieren lässt, dass der Bewerber Beschreibungen mehrerer realistischer Situationen vorgelegt respektive Filmsequenzen vorgespielt bekommt, welche er in der zu besetzenden Stelle mit hoher Wahrscheinlichkeit antreffen wird. Bei jeder Situation muss der Bewerber anhand einer Liste möglicher Handlungsalternativen angeben, wie er sich in dieser Situation am ehesten verhalten würde, respektive wie in der geschilderten Situation am effektivsten gehandelt werden sollte. Die Antworten werden dann auf Grund ihres relativen Effektivitätslevels bewertet (z. B. Chan & Schmitt, 1997, 2002; Hanson, Horgen & Borman, 1998; Lievens, Peeters & Schollaert, 2008; McDaniel, Morgeson, Finnegan, Campion & Braverman, 2001; McDaniel & Nguyen, 2001; Weekly & Jones, 1997, 1999; Weekley & Ployhart, 2006). SJTs sind relativ einfach und kostengünstig zu entwickeln (Clevenger, Pereira, Wiechmann, Schmitt & Schmidt Harvey, 2001) und werden als Ersatz für aufwändigere Simulationsverfahren – zum Beispiel Assessment Centers – eingesetzt (Motowidlo, Dunnette & Carter, 1990).

Zu Beginn noch als Konstrukt aufgefasst (z. B. Sternberg, Wagner, Williams & Horwath, 1995) werden heute SJTs als eine Messmethode angesehen, mit welchen sich eine Vielzahl von Konstrukten erfassen lassen (z. B. Ployhart & Ehrhart, 2003; Weekley & Jones, 1997; ausführliche Darstellung der Methoden-Konstrukt-Diskussion in Schmitt & Chan, 2006), welche Arbeitsleistung valide messen (z. B. Chan & Schmitt, 2002; McDaniel et al., 2001; Motowidlo et al., 1990), über eine hohe Augenscheinvalidität verfügen (z. B. Rosen, 1961; Van Vianen, Taris, Scholten & Schinkel, 2004) und inkrementelle Validität gegenüber anderen Selektionsverfahren aufweisen (z. B. Chan & Schmitt, 2002; Clevenger et al., 2001; McDaniel, Powell Yost, Ludwick, Hense & Hartmann, 2004; O'Connell, Hartman, McDaniel, Grubb & Lawrence, 2007; Weekley & Jones, 1997, 1999).

Bei SJTs handelt sich somit um eine für ein spezifisches Tätigkeitsgebiet entwickelte, indirekte Form der Arbeitssimulation, welche jedoch nicht eine oder

mehrere vordefinierte Verhaltens- oder Persönlichkeitsdimensionen erfassen sondern multidimensionale Verfahren darstellen, welche arbeitsbezogenes Wissen und Verhalten, interpersonale Fähigkeiten, Team- und Führungsverhalten und verschiedene Persönlichkeitsdimensionen erfassen und welche hauptsächlich bei der Auswahl von Führungskräften zum Einsatz gelangen (Christian, Edwards & Bradley, 2010; McDaniel et al., 2001; Motowidlow et al., 1990). Einige Autoren sind der Ansicht, dass SJTs praktische Intelligenz im Sinne von Wagner und Sternberg (1985) erfassen, also die Fähigkeit, effektiv und erfolgreich auf Anforderungen in unterschiedlichsten Situationen zu reagieren (z. B. Chan & Schmitt, 2002; Motowidlo et al. 1990; Sternberg, 1998, 1999), wobei Schmidt und Hunter (1993) die Meinung vertreten, dass es sich bei der praktischen Intelligenz oder dem Erfahrungswissen (*tacit knowledge*) um nichts anderes als Berufswissen handelt. Da SJTs jedoch nicht einen generellen Faktor erfassen – auch nicht praktische Intelligenz – und SJTs zu Intelligenz inkrementelle Validität bezüglich Berufserfolg aufweisen sind McDaniel und Whetzel (2005) der Ansicht, dass SJTs als Messmethode und nicht als Konstrukt zu verstehen sind.

SJTs werden zu den situationalen Testverfahren gezählt, welche auch Assessment Centers (Thornton & Byham, 1982), Situationale Interviews (Latham, Saari, Pursell & Campion, 1980) oder Videosimulationen (Weekley & Jones, 1997) umfassen. Gemeinsames Merkmal der Verfahren ist, dass die Personalfachleute den Bewerbern Situationen aus dem Arbeitsalltag präsentieren und ihre Reaktion darauf erfassen respektive beobachten und beurteilen. Im Gegensatz zur Arbeitsprobe oder zum Assessment-Center müssen die Bewerber beim SJT jedoch nicht in einer realen oder realistischen Arbeitssituation konkretes Verhalten zeigen, sondern haben aufgrund der Beschreibung einer hypothetische Arbeitssituation lediglich Verhaltensabsichten zu schildern. Dies bezeichneten Motowidlo et al. (1990) als *low fidelity simulations*, ersteres als *high fidelity simulations*. Mehrere Autoren führen als einen der Gründe für das steigende Interesse an SJTs (Ployhart, 2006; Salgado, Viswesvaran & Ones, 2001) den hohen finanziellen, organisatorischen und personellen Aufwand bei der Durchführung von Realsimulationen und Interviews an (Clevenger et al., 2001; Motowidlo et al., 1990; Weekley & Jones, 1999). Da die Kosten bei der Entwicklung und Druchführung von SJTs relativ tief ausfallen, eigenen sie sich besonders für eine erste Triage bei grossen Bewerbergruppen (*select out*), wohingegen Assessment Centers oder Situationale Interviews in der Endphase des Selektionsprozesses bei einer kleinen, ausgewählten Gruppe aus dem Bewerberpool (*select in*) eingesetzt werden (Lievens et al., 2008). Als weiteren Grund für das gestiegene Interesse nennen Weekley und Jones (1999) und Hanson et al. (1998) die Möglichkeit einer alternativen Selektionsmethode mit guter Validität (Meta-

Analyse von McDaniel et al., 2001: ρ = .34) jedoch höherer Test-Fairness (geringerer *adverse impact*) als zum Beispiel bei Intelligenztests (Chan & Schmitt, 2002; Chan, Schmitt, DeShon, Clause & Delbridge, 1997; Clevenger et al., 2001; Hanson & Borman, 1995; Motowidlo et al., 1990; Olson-Buchanan, Drasgow, Moberg, Mead, Keenan & Donovan, 1998; Weekley & Jones, 1997, 1999; Weekley, Ployhart & Harold, 2004). Eine von Whetzel, McDaniel und Nguyen (2008) durchgeführte Meta-Analyse mit SJTs ergab zwar Rassenunterschiede, welche sich jedoch zu einem grossen Teil mit Intelligenzunterschieden erklären lassen. In mehreren Studien liess sich auch nachweisen, dass Bewerber SJTs in etwa gleich gut akzeptieren wie ein situatives Interview (Banki & Latham, 2010). Bei der Beurteilung des Tätigkeitsbezuges (*job relatedness*) gibt es keinen Unterschied zwischen einer text- und einer video- respektive computerbasierten Version eines SJTs (Kanning, Grewe, Hollenberg & Hadouch, 2006), hingegen bei der Augenscheinvalidität (Chan & Schmitt, 1997; Richman-Hirsch, Olson-Buchanan & Drasgow, 2000). Eine Übersicht zum Thema Bewerberreaktionen bei SJTs liefern Bauer und Truxillo (2006).

McDaniel und Whetzel (2007) zeigen die Vielfalt der Einsatzmöglichkeiten von SJTs anhand einer Liste mit nach dieser Methode entwickelten Testverfahren auf, welche zum Beispiel Teamleistung (Stevens & Campion, 1999), Teamrolle (Mumford, Van Iddekinge, Morgeson & Campion, 2008), Integrität (Becker, 2005), Verkaufserfolg (Phillips, 1992), Verhalten bei Verhandlungen (Phillips, 1993), initiatives Verhalten (Bledow & Frese, 2009), die Vorhersage der Kündigung bei Versicherungsvertretern (Dalessio, 1994), den Berufserfolg bei Ingenieuren (Clevenger, Jockin, Morris & Anselmi, 1999) oder die Leistung im College (Oswald, Schmitt, Kim, Ramsay & Gillespie, 2004) erfassen.

SJTs basieren auf der Überlegung, dass in der Vergangenheit gezeigtes Verhalten den besten Prädiktor für zukünftiges Verhalten darstellt (*behavioral consistency principle*; Wernimont & Campbell, 1968; siehe auch Motowidlo et al., 1990). Typischerweise sind SJTs als Multiple-Choice-Tests konzipiert und werden als Paper-&-Pencil-Verfahren, computerisiert oder in Video-Szenarien vorgelegt. Nach McDaniel et al. (2001) zählen auch diejenigen Verfahren zur Gruppe der SJTs, welche anstelle von Situationsbeschreibungen auf Statements basieren, die der Bewerber hinsichtlich deren arbeitsbezogener Angemessenheit einstufen muss. Die verschiedenen SJTs lassen sich hauptsächlich bezüglich der Ausprägung der Spezifität der geschilderten Arbeitssituationen unterscheiden: Einen Berufseinsteiger lässt man eher allgemein gehaltene Arbeitssituationen beurteilen – zum Beispiel Situationen zur Arbeitsorganisation oder zum Teamwork – währenddem man für die Besetzung einer Führungs– oder einer Fachposition Situationen vorlegt, deren erfolgreiche Bewältigung respektive korrekte Einschät-

zung grosse Erfahrung oder fachspezifisches Wissen verlangen (McDaniel & Whetzel, 2007).

Seit den Beiträgen von Wagner und Sternberg (1985) zum *tacit knowledge* und von Motowidlo, Dunnette und Carter (1990) zur *low fidelity simulation* wird den SJTs ein immer grösser werdendes Forschungsinteresse zuteil, was zu einer grossen Anzahl wissenschaftlicher Studien geführt hat. Mittlerweile beschreiben Autoren SJTs auch in Übersichtswerken zu Human Ressource Management und Personalselektion (z. B. Chan & Schmitt, 2005; McDaniel & Whetzel, 2007) und Weekley und Ployhart gaben 2006 die erste Monografie dazu heraus. Die Aktualität der SJTs darf nicht darüber hinwegtäuschen, dass diese Verfahren schon seit langer Zeit zum Einsatz gelangen: Als erster SJT gilt der „George Washington Social Intelligence Test" (Moss, Hunt, Omwake & Ronning, 1927), in welchem die Subskala *Judgment in Social Situations* enthalten ist (Kihlstrom & Cantor, 2000; McDaniel et al., 2001). Als weitere Beispiele für frühe SJTs nennen Motowidlo et al. (1990) unter anderem die Papier-und-Bleistift-Tests „Practical Intelligence Test" (Cardall, 1942), den „Business Intelligence Test" (Bruce, 1965), „How Supervise?" (File, 1943) oder den „Supervisory Practices Test" (Bruce & Learner, 1958). Vor allem das Militär setzte schon früh SJTs in der Selektion ein (Hanson, Horgen & Borman, 1998; McDaniel & Whetzel, 2005): Während des zweiten Weltkrieges waren mehrere SJTs in der U.S. Army im Einsatz (Guilford & Lacey, 1947). Abblidung 2.1 zeigt ein Item-Beispiel aus der ersten Version der „Army Air Force Qualifying Examination" aus dem Jahre 1942.

A pilot has to make a forced landing near a mountain cabin. He finds that the nearest phone is at an isolated fire ranger's cabin 14 miles across the mountains to the north. It is winter. He sets out on foot for the ranger's cabin at 6 a.m., carrying food for only one meal. At 10 a.m., having met no one, he comes to three branches of the trail, all unmarked. His most practical decision would be to:

A. Follow the trail which appears to lead in the right direction until he reaches the cabin or the end of the trail.

B. Turn back immediately to his starting point.

C. Leave the trail and go due north by compass.

D. Walk along the trail which appears to lead in the right direction until noon, then turn back if not sure of his location.

E. Stay in the fork in the trail and wait for someone to come by.

Abbildung 2.1 Beispiel eines SJT-Items aus der „Army Air Force Qualifying Examination" (Guilford & Lacey, 1947, S. 124).

2.2 Die Konstruktion von Situational Judgment Tests

Die Verwendung von realen Alltagssituationen als Kern von SJT-Items erfordert ein spezielles Vorgehen bei der Testkonstruktion, welches sich deutlich von demjenigen bei herkömmlichen Persönlichkeitsinventaren unterscheidet. Motowidlo et al. (1990; siehe auch McDaniel & Nguyen, 2001) beschrieben folgendes – als klassisch zu bezeichnendes – Vorgehen für die Konstruktion von SJTs:

1. *Subject Matter Experts* (SMEs) formulieren Critical Incidents in Arbeitssituationen (*domain sampling approach*).
2. Eine zweite Gruppe von SMEs gibt zu jeder Situation an, wie sie sich darin verhalten würde.
3. Anhand dieses Materials erstellen die Testkonstrukteure die erste Version des SJTs.
4. Eine dritte Gruppe von SMEs überarbeitet die Items und gibt an, welche der Verhaltensalternativen die beste und welche die schlechteste ist.

Das in den letzten fünfzehn Jahren verstärkte Forschungsinteresse an SJTs hat dazu geführt, dass Forscher in der aktuellen Literatur das Vorgehen bei deren Entwicklung viel differenzierter beschreiben. Die von Weekley, Ployhart und Holtz (2006) aufgeführten Tätigkeiten lassen sich in zwei grosse Konstruktionsschritte aufteilen: Die Entwicklung des Testmaterials (Situation, Antwortalternativen und Instruktion) und die Entwicklung der Auswertungsmethode und des Bewertungsschlüssels (Effektivität der Antwortalternativen und Scoring-Methode). McDaniel, Whetzel und Nguyen (2006; siehe auch McDaniel & Nguyen, 2001; McDaniel & Whetzel, 2007; Weekley et al., 2006) führen sieben Schritte bei der Entwicklung eines SJTs auf:

1. Wahl der Arbeitsstellen respektive Tätigkeiten, für welche der SJT entwickelt werden soll und Eingrenzung des Testinhaltes.
2. Sammlung erfolgskritischer Verhaltensweisen (Critical Incidents).
3. Sortieren der Critical Incidents.
4. Bildung des Item-Stamms anhand der Critical Incidents.
5. Generierung der Antwortalternativen.
6. Wahl der Antwortinstruktionen.
7. Entwicklung des Auswertungsschlüssels.

Nachfolgend gehe ich auf die einzelnen Schritte detailliert ein.

Schritt 1: Wahl der Tätigkeitsbereiche oder Positionen, für welche der SJT entwickelt werden soll und Eingrenzung des Testinhaltes

Bevor die Testentwickler mit ihrer Arbeit beginnen können, ist genau zu definieren, für welche Tätigkeitsbereiche oder Positionen innerhalb der Organisation der SJT zum Einsatz gelangen soll. Zur genauen Beschreibung dieser Bereiche oder Positionen können die klassischen Arbeitsanalyseverfahren eingesetzt werden. Je enger sich dabei der Arbeitsinhalt charakterisieren lässt, desto spezifischer wird der SJT ausfallen. Die Auftraggeber oder die Testentwickler müssen auch entscheiden, ob technisches Wissen im Test enthalten sein soll. Vor allem bei Arbeitsstellen mit schnellem Technologiewandel ist zu berücksichtigen, dass der Testinhalt unter Umständen nach wenigen Jahren veraltet ist.

Schritt 2: Sammlung erfolgskritischer Verhaltensweisen (Critical Incidents)

Weekley et al. (2006) beschreiben auf Grund ihrer Literaturrecherche zwei Methoden zur Auswahl der in einem SJT verwendeten Situationen (als Item-Stamm bezeichnet): Die *Critical Incidents Technique* (induktives Vorgehen) und *theorie-basierte Methoden* (deduktives Vorgehen).

Als häufigste Methode setzen die Testentwickler die Critical Incidents Technique (Flanagan, 1954; siehe auch Anderson & Wilson, 1997; Butterfield, Borgen, Amundson & Maglio, 2005) ein. Dabei beschreiben SMEs – häufig Stelleninhaber und/oder Vorgesetzte, aber auch Kunden – erfolgreiche und ineffektive Verhaltensweisen innerhalb eines Arbeitsgebietes. Diese werden in Workshops gesammelt, in welchen die Durchführungsverantwortlichen die SMEs bitten, von Verhaltensweisen, die sie selbst ausgeführt haben oder die sie bei anderen Stelleninhabern beobachten konnten die Ausgangssituation, die konkret vom Stelleninhaber unternommene Handlung und das Ergebnis zu schildern. McDaniel et al. (2006) lassen die SMEs zusätzlich noch angeben, welche Kompetenz des Stelleninhabers für das beschriebene Ereignis bedeutsam ist. Anderson und Wilson (1997, siehe auch McDaniel et al., 2006) führen eine Liste mit Anweisungen auf, welche den Experten als Hilfestellung für die Suche nach erfolgskritischen Verhaltensweisen gegeben werden können (z. B. „Denken Sie an ein Ereignis, bei dem jemand eine wirklich gute Leistung zeigte.", „Denken Sie an eine Person bei der Arbeit, welche Sie bewundern. Können Sie sich an ein Ereignis erinnern, bei welchem sich diese Person als aussergewöhnlich leistungsfähig bewies?" oder „Denken Sie an Fehler bei der Arbeitserledigung, welche Sie bei neu eintretenden Stelleninhabern festgestellt haben."). Latham und Wexley (1982) empfehlen, dass mindestens 300 Ereignisse von mindestens 30 Experten für eine Testentwicklung zu sammeln sind, was McHenry und Schmitt (1994) als absolu-

tes Minimum bezeichnen und auf Motowidlo et al. (1990) verweisen, welche 1200 Ereignisse von 139 Experten gesammelt haben, um ihren *Work Sample Test* für Manager im Telekommunikationsbereich zu entwickeln. McDaniel et al. (2006) geben an, dass ein SME innerhalb von zwei Stunden fünf bis zehn Critical Incidents beschreiben kann.

Neben der Critical Incidents Technique zur Bildung des Ausgangsmaterials für die Itementwicklung gelangen – wie weiter oben schon erwähnt – auch theorie-basierte Methoden zum Einsatz. So hat zum Beispiel Becker (2005) zu den sieben Dimensionen seines Modells zu integerem Verhalten in Organisationen (Becker, 1998, 2000) theoriegeleitet 30 Items formuliert. Dasselbe Vorgehen haben Stevens und Campion (1999) für die Entwicklung ihres Teamwork-Tests gewählt. Chan (Chan & Schmitt, 2002) und Reynolds, Winter und Scott (1999) haben ein gemischtes Verfahren eingesetzt: Anhand von Arbeitsplatzanalysen, Interviews und Untersuchungen haben sie relevante Dimensionen definiert, zu welchen sie Stelleninhaber und Manager erfolgskritische Verhaltensweisen formulieren liessen, wobei Chan nur Situationen in den Test aufgenommen hat, welche eindeutig verhaltens- und situationsbezogene Anforderungen beschreiben. Ähnlich sind Oswald et al. (2004) bei der Entwicklung eines SJTs für die Vorhersage der Leistung von College-Studenten vorgegangen, indem sie in einem ersten Schritt in den Internetauftritten von Colleges und Universitäten nach deren Lernzielen oder Leitbildern suchten. Die daraus resultierenden 174 Aussagen von 34 Institutionen liessen sie von Experten zu Clustern sortieren. In einer Gruppendiskussion einigten sich diese schliesslich auf zwölf relevante Dimensionen. Die Autoren formulierten daraufhin SJT-Items von bestehenden Testverfahren um, damit sie zu einer der zwölf Dimensionen passen und liessen Studenten Critical Incidents zu den Dimensionen beschreiben um daraus weitere Item-Stämme zu formulieren.

McDaniel et al. (2006) geben an, dass sich mit einem solchen Vorgehen eine erste Version des SJTs an einem Tagesworkshop erstellen lässt (Schritte 2 bis 5). Weekley et al. (2006) raten, Verhaltensdimensionen, welche aus gründlich durchgeführten Arbeitsplatzanalysen abgeleitet sind und sich als wichtig für die Arbeitsleistung erweisen, in den Testentwicklungsprozess einzubeziehen. Warnend fügen sie jedoch hinzu:

> It is important to note that although numerous attempts have been made to develop SJTs around constructs or competencies, few of these developers have sought to create subscores reflecting these dimensions. Probably due to the multidimensionality and poor psychometric characteristics of the subscales, most have instead collapsed across items to create an

> overall SJT score. ... Furthermore, as SJT development methods become more refined, it may be possible to develop SJTs that capture unique dimensional constructs. At present, developing a model, from theory or job analyses, and then collecting critical incidents from a variety of sources to fit the model, would appear to be the most comprehensive means of developing SJT stems. (Weekley et al., 2006, S. 160-161)

Schritt 3: Sortieren der Critical Incidents

Im Anschluss an die Sammlung der Critical Incidents gruppieren die Testentwickler diese nach den darin geschilderten Situationen. Idealerweise sind die mit dieser Arbeit beauftragten Personen mit den Inhalten und Tätigkeiten der im SJT abgedeckten Arbeitsstelle respektive Arbeitsstellen vertraut. In diesem Schritt sind mehrere Ziele zu erreichen:

- Identifikation von gleichen oder ähnlichen Situationsschilderungen.
- Erkennen von nicht abgedeckten Arbeitssituationen.
- Bestimmen der Bereiche, zu welchen die Item-Stämme geschrieben werden.
- Identifikation von ungeeigneten Inhalten, zum Beispiel solchen, die man Bewerbern nicht vorlegen möchte, wie Diskrimination, fehlende Aufstiegsmöglichkeiten oder die Einführung unpopulärer Massnahmen. Die Testentwickler sollten hier auf jeden Fall Entscheidungsträger der Organisation einbeziehen, damit auch diese die Möglichkeit haben, aus ihrer Sicht ungeeignete Verhaltensweisen auszuschliessen.

McDaniel et al. (2006) führen eine Liste mit Inhaltsbereichen auf, welche die SMEs typischerweise im Rahmen der Critical Incidents Technique nennen und welche sich in die zwei Bereiche Arbeit und Arbeitsinhalt (z. B. zu viel Arbeit, herausfordernde Arbeit) und Probleme mit Personen am Arbeitsplatz (z. B. Kollegen, Unterstellten, Chefs) gliedern lassen.

Schritt 4: Bildung des Item-Stammes anhand der Critical Incidents

Die Testentwickler stehen nun vor der sehr zeitintensiven Aufgabe, das gesammelte und sortierte Ausgangsmaterial zu überarbeiten und in eine zuvor definierte sprachliche Form und Länge zu bringen. Hier ist besonders darauf zu achten, dass die Situationen klar und gut verständlich beschrieben sind und man ein einheitliches Vokabular einsetzt, indem man zum Beispiel konsequent den Term Vorgesetzter (vs. Chef oder Boss) verwendet. In diesem Schritt gilt es auch zu

entscheiden, wie viele einzelne erfolgskritische Situationen zu einem bestimmten Thema, zum Beispiel zur Zusammenarbeit mit dem Vorgesetzten, im Test enthalten sein sollen, respektive wie breit ein einzelnes Thema erfasst werden soll. Zudem muss man ein besonderes Augenmerk darauf legen, dass die beschriebene Situation für alle Arbeitsstellen, für welche der SJT zum Einsatz gelangen soll, Gültigkeit besitzt und bis zu einem gewissen Grad repräsentativ oder prototypisch ist. Es ist auch gut möglich, dass einige der Bewerber gewisse der geschilderten Situationen aus dem Berufsalltag noch nie erlebt haben. Hanson et al. (1998) schlagen deshalb vor, dass ausreichend verallgemeinerte Situationen zu generieren sind, so dass die meisten Bewerber über genügend Erfahrungen verfügen, um die dazu formulierten Verhaltensalternativen auch bezüglich deren Effektivität einschätzen zu können.

A costumer asks for a specific brand of merchandise the store doesn't carry. How would you respond to the costumer?

1) Tell the costumer, which stores carry that brand, but point out that your brand is similar.
2) Ask the costumer more questions so you can suggest something else.
3) Tell the costumer that the store carries the highest quality merchandise available.
4) Ask another associate to help.
5) Tell the customer which stores carry that brand.

Your organization purchased an applicant-tracking software package a couple of months back and it's finally up and running. You have just been assigned to coordinate the team responsible for training recruiters to use the program. The recruiting season starts in 3 weeks and your supervisor has informed you that all recruiters must be trained and proficient in the new software before the season starts.

One of the trainers assigned to your team informs you that he is scheduled for 7 days of medical leave starting next week. The trainer scheduled the medical leave months ago to allow recovery time for a minor surgical procedure. Rate the effectiveness of the following responses.

1) Ask your supervisor for another trainer to replace the individual going on leave.
2) Inform the trainer that he is responsible for finding a co-worker to fill in in his absence.
3) Inform your supervisor that you will not be able to train all recruiters before the season starts because you are short one trainer for an entire week.

Abbildung 2.2 Beispiele von niedrig- und hochkomplexen Item-Stämmen (Weekley et al., 2006, S. 162–163).

Die Item-Stämme lassen sich auf Grund der Komplexität respektive Differenziertheit der geschilderten Situation in *low*- und *high-complexity* SJT-Items einteilen (McDaniel & Nguyen, 2001; siehe auch Abbildung 2.2). Bei hochkomplexen Item-Stämmen, welche durch eine ausführliche Situationsbeschreibung gekenntzeichnet sind, ist zu beachten, dass diese zu Subgruppendifferenzen auf Grund der unterschiedlichen sprachlichen Kompetenzen der Testbearbeiter führen können (Sacco, Schmidt & Rogg, 2000). Empirisch noch nicht eindeutig belegen liess sich die Auswirkung der Komplexität des Item-Stamms auf die Validität: Anzunehmen wäre, dass SJT-Items mit im Detail geschilderten Arbeitssituationen bessere Prädiktoren für tatsächliches Verhalten darstellen als rudimentäre, vereinfachte Schilderungen, da die hohe Übereinstimmung zwischen dem Test- und dem Arbeitsinhalt als eine Erklärung für die Validität der SJT angesehen wird (Chan & Schmitt, 2002; Lievens, Buyse & Sackett, 2005). So zeigten Reynolds et al. (1999) in ihrer Studie auf, dass spezifisch formulierte Items eine höhere Validität aufweisen als allgemein formulierte. McDaniel et al. (2001) stellten hingegen in ihrer Meta-Analyse fest, dass die Kriteriumsvalidität bei hochkomplexen SJT-Items tiefer ausfällt als bei allgemein formulierten. Einen Vorteil detailliert und spezifisch formulierter Items orten Weekley et al. (2006) in der höheren Augenscheinvalidität und der damit verbundenen besseren Reaktionen der Bewerber auf den Test.

Schritt 5: Generierung der Antwortalternativen

Die Antwort- respektive Verhaltensalternativen zu den einzelnen Item-Stämmen sollen mehrere plausible Möglichkeiten aufzeigen, wie Menschen auf Grund ihrer Erfahrung, ihres Vorwissens oder ihrer Persönlichkeit in der geschilderten Situation reagieren. Üblicherweise werden dem Testbearbeiter zu jedem Item-Stamm zwischen drei und zwölf mögliche Verhaltensweisen präsentiert (Weekley et al., 2006). Deren Generierung übernehmen entweder die Testentwickler selbst (z. B. Stevens & Campion, 1999; Weekley & Jones, 1999) oder es gelangen erneut Experten zum Einsatz (z. B. Hunter, 2003; Motowidlo, Hanson & Crafts, 1997). Letzteres Vorgehen hat den Vorteil, dass die SMEs mehr Alternativen erarbeiten und die Antworten realistischer ausfallen, da zum Beispiel auch fachspezifisches Wissen einfliessen kann. Da ein SME in der Regel nur etwa zwei bis drei alternative Verhaltensweisen schildern kann, müssen dieselben Item-Stämme mehreren SMEs vorgelegt werden. Um den SMEs ihre Arbeit zu erleichtern, geben ihnen McDaniel et al. (2006) folgende Hinweise für die Generierung der Verhaltensalternativen:

- Was würden Sie in dieser Situation tun?
- Was ist das erfolgversprechendste Verhalten?
- Was würde ein effizienter (ineffizienter) Arbeitnehmer tun?
- Denken Sie an einen guten (ungenügenden) Arbeitskollegen. Was würde dieser tun?

Die Testentwickler müssen die auf diese Weise generierten Antworten sprachlich überarbeiten und sich für eine – zuvor festgelegte – Anzahl davon entscheiden. Dabei ist zu berücksichtigen, dass man keine Verhaltensweisen wählt, für die sich keiner oder praktisch jeder der zukünftigen Testbearbeiter entscheidet.

Neben dieser als *expertenbasiert-empirisch* zu bezeichnenden Vorgehensweise zur Generierung der Antwortalternativen führen Weekley et al. (2006) noch eine von ihnen als *construct-based* beschriebene auf. Im Gegensatz zu den meisten der expertenbasiert-empirisch konstruierten SJT, bei welchen alle Antworten zu einem Index aufsummiert werden, repräsentieren die Verhaltensalternativen bei den konstruktbasierten SJTs Indikatoren spezifischer Konstrukte, welche in Subskalen gruppiert zu mehreren Summenscores verrechnet werden. Als Beispiele seien an dieser Stelle die SJTs zur Erfassung von Big Five-Dimensionen genannt, welche Trippe und Foti (2003; siehe auch Trippe, 2002) und Motowidlo, Diesch und Jackson (2003; siehe auch Motowidlo, Hooper & Jackson, 2006b) entwickelten. Im Hinblick auf die später beschriebene eigene Testentwicklung gehe ich an dieser Stelle vertieft auf konstruktbasierte SJT-Skalen ein.

It is the beginning of the semester and your professor notes that a significant portion of your grade will be based on a fairly comprehensive project that will be due the last day of class. *What would you do?*

1) Start working on the project right away.
2) Start working on the project at mid term.
3) Start working on the project near the end of the semester.
4) Start working on the project a few days before it is due.

Abbildung 2.3 Beispiel eines Big Five-SJT-Items zur Dimension Gewissenhaftigkeit (Trippe, 2002, S. 64).

Trippe und Foti (2003) operationalisierten die Dimensionen Gewissenhaftigkeit, Verträglichkeit und Offenheit für Erfahrung, indem sie zu Items aus dem IPIP

(International Personality Item Pool; http://ipip.ori.org; siehe auch Goldberg, 1999; Goldberg et al., 2006) Situationen beschrieben, zu welchen sie vier bis fünf Verhaltensalternativen entwickelten. „That is, each item was directly linked to a specific personality trait by writing responses to reflect varying degrees of the trait described in each of the corresponding items from the IPIP" (Trippe, 2002, S. 24). Das in Abbildung 2.3 dargestellte SJT-Item basiert auf dem IPIP-Gewissenhaftigkeits-Item „start tasks right away". Trippe formulierte die vier Verhaltensalternativen „on a continuous (1-4) scale" (Trippe, 2002, S. 24) so, dass die Verhaltensalternative 1 eine hohe Ausprägung in der Dimension Gewissenhaftigkeit und die Verhaltensalternative 4 eine tiefe Ausprägung repräsentiert.

Insgesamt hat Trippe 3 x 15 Items formuliert, welche er zur Kontrolle von 14 Studenten einer der drei Dimensionen zuordnen liess. Es erreichten 18 der 45 Items das Kriterium von 86% Übereinstimmung, wobei er 3 x 8 Items in den Fragebogen aufnahm. Für die Berechnungen hat Trippe jedoch nur jeweils die vier Items mit den höchsten Interkorrelationen pro Dimension berücksichtigt. Die Analysen ergaben, dass der SJT diskriminante und konvergente Validität aufweist, jedoch im Vergleich zu den IPIP-Items ein signifikanter Methodenfaktor auftritt, in dem Sinne, dass im Modell die Trait-Ladungen kleiner sind als die Methoden-Ladungen. Dies ist problematisch, da eine Verhaltensvorhersage auf Grund dieser Methodenvarianz, welche unabhängig von der zu messenden Persönlichkeitseigenschaft ist, ungenau oder sogar falsch ausfallen kann (Messick, 1995). Trippe (2002, S. 35) gelangt auf Grund seiner Studie zu folgendem Schluss: „With the proper developmental rigor, it seems reasonable that similar procedures for constructing narrowly focused SJT scales can generalize to more realistic SJT's designed to measure traditional performance dimensions."

In ihrer Meta-Analyse zur Validität von Assessment Centern stellten Lievens und Conway (2001) fest, dass sich der Einfluss der Dimensionsfaktoren im Vergleich zu den Übungsfaktoren vergrössern lässt, wenn weniger Dimensionen erhoben und anstelle von Managern Psychologen als Assessoren eingesetzt werden, da sich letztere im Studium mit der Erfassung und Beschreibung individueller Unterschiede auseinander gesetzt haben. Übertragen auf die Konstruktion von SJTs schliesst Trippe (2002, S. 32):

> SJT's often use job incumbents and non-psychologist SME's to develop item vignettes and stems. It is reasonable to believe that reliance on individuals who do not have extensive knowledge of individual difference variables or psychometric principles to develop dimension scales will result in a weaker measure with regard to convergent and discriminant validity.

Motowidlo et al. (2003) entwickelten ausgehend von der *implicit trait policy* (ITP; Motowidlo, Hooper & Jackson, 2006a, 2006b) einen SJT zur Erfassung einzelner Big Five-Dimensionen. Die ITP beschreibt implizite Annahmen über den kausalen Zusammenhang zwischen Persönlichkeitszügen und effektivem Verhalten und liefert so eine Erklärung für den Zusammenhang zwischen den Messwerten von Persönlichkeitstests und SJTs. (Eine detailliertere Beschreibung der in der ITP getroffenen Annahmen folgt in Kapitel 2.3.) Sie entwickelten insgesamt 16 Item-Stämme, welche sich auf die Dimensionen Verträglichkeit, Extraversion und Gewissenhaftigkeit beziehen. Dazu formulierten sie pro Item-Stamm je fünf bis zehn Verhaltensalternativen, wovon je die Hälfte einer hohen Ausprägung in der entsprechenden Dimension und die andere Hälfte einer tiefen Ausprägung entspricht. Die Probanden erhielten den Auftrag, jede Verhaltensalternative auf einer siebenstufigen Anwortskala, welche von sehr ineffektiv bis sehr effektiv reicht, einzustufen. Motowidlow et al. bildeten die Skalenscores zu den drei Persönlichkeitsdimensionen, indem sie von den Einstufungen zu den Verhaltensalternativen mit der hohen Trait-Ausprägung diejenigen zu den niedrigen subtrahierten. Die Korrelationen zu den mit dem NEO-FFI gemessenen Dimensionen betragen .32 (.31) für Verträglichkeit, .34 für Extraversion und .17 (-.06) für Gewissenhaftigkeit (die Werte in Klammern beziehen sich auf eine zweite Version mit je 19 Items). Motowidlo et al. (2006a, 2006b) interpretieren diese Ergebnisse dahingehend, dass ITP-Scores in Übereinstimmung mit der von ihnen formulierten Theorie implizite Messungen von Persönlichkeitsdimensionen darstellen.

Ployhart und Ryan (2000; siehe auch Porr & Ployhart, 2004; Ployhart, Porr & Ryan, o. J.) stützten sich bei ihrem Versuch, mit einem SJT Persönlichkeitskonstrukte zu erfassen, auf die *competency-demand hypothesis* (Mischel & Shoda, 1995; Shoda, Mischel & Wright, 1993), welche besagt, dass eine Situation, deren Bewältigung eine hohe Ausprägung in einer spezifischen Persönlichkeitsdimension verlangt, eine grosse Variabilität in der Manifestation eben dieser aufweisen wird. Indem Wright und Mischel (1987) aufzeigten, wie sich Situationen bezüglich der zur Bewältigung benötigten Fähigkeiten oder Persönlichkeitsdimensionen unterscheiden lassen, liefern sie die theoretische Grundlage für die explizite Verknüpfung von in SJT geschilderten Situationen mit spezifischen Persönlichkeitsdimensionen. Ployhart et al. (o. J.) wählten bei der Entwicklung des SJT eine konstruktorientierte Vorgehensweise, indem sie zu den für im Umgang mit Kunden wichtigen Persönlichkeitsdimensionen Neurotizismus, Verträglichkeit und Gewissenhaftigkeit Item-Stämme entwickelten.

In einem ersten Schritt definierten sie anhand eines Literaturstudiums die Leistungsdimensionen für Kundenkontakt. Weiter führten sie eine Arbeitsanalyse

durch, indem sie Studenten, welche Arbeiten mit Kundenkontakt ausübten, typisches Verhalten beschreiben liessen. Im zweiten Schritt liessen sie sich Situationen im Kontakt mit Kunden schildern, welche häufig auftreten, schwierig zu bewältigen und wichtig sind und somit eine maximale Variabilität im gezeigten Verhalten erwarten lassen. Im nachfolgenden Konstruktionsschritt ordneten Experten die gesammelten Situationen den Dimensionen aus der Arbeitsanalyse zu. Im vierten Schritt formulierten sie anhand eines Konstruktionsrationals Verhaltensalternativen für jede Situation. Dieses führt zu Verhaltensalternativen, welche sich auf einem Verhaltenskontinuum von einer sehr geringen Ausprägung auf der zu messenden Dimension bis zu einer hohen einstufen lassen (siehe Abbildung 2.4). Im fünften Schritt stuften studentische Experten die Relevanz und die Angemessenheit der Situationen und Verhaltensalternativen ein und bestimmten deren Zusammenhänge zu Neurotizismus, Verträglichkeit und Gewissenhaftigkeit. In den definitiven Fragebogen nahmen die Autoren nur diejenigen Situationen auf, welche die Experten eindeutig einer Persönlichkeitsdimension zugeordnet haben. Es zeigte sich, dass die Scores der SJT-Dimensionen mässig mit den entsprechenden Werten aus dem NEO-FFI korrelierten ($r_{Neurotizismus}$ = -.10, $r_{Verträglichkeit}$ = .17, $r_{Gewissenhaftigkeit}$ = .34) und tief mit Intelligenz ($r_{Neurotizismus}$ = .18, $r_{Verträglichkeit}$ = .04, $r_{Gewissenhaftigkeit}$ = -.03). Als potenziellen Nachteil dieses Itemformates sehen Ployhart et al. eine einfachere bewusste Verfälschung (*faking*) der Antworten im Vergleich zu einem klassischen SJT, bei welchem bei der Konstruktion darauf geachtet wurde, dass die Antwortaltenativen ähnlich sozial erwünscht sind.

A customer is telling you a personal story that is completely unbelievable and that you think is factually incorrect. What do you do?

A. Nod in agreement with the customer, then correct the customer's story.

B. Nod in agreement with the customer, but say nothing.

C. Smile an nod in agreement with the customer, but say nothing.

D. Smile an nod in agreement with the customer, and say you agree with the story.

Abbildung 2.4 Beispiel eines SJT-Items zur Erfassung der Servicequalität (Ployhart et al., o. J., S. 42).

Stemler und Sternberg (2006) haben einen SJT zur Erfassung von praktischer Intelligenz bei Lehrpersonen entwickelt. Dazu haben sie anhand von Interviewmaterial Item-Stämme gebildet, zu welchen sie jeweils Verhaltensalternativen zu sieben, empirisch hergeleiteten Strategien im zwischenmenschlichen Umgang

von Lehrpersonen formulierten (Stemler, Elliot, Grigorenko & Sternberg, 2006). Ausgegangen sind sie von einem Modell erfolgreicher Intelligenz (*successful intelligence*), in welchem sich kreative, analytische und praktische Intelligenz gegenseitig beeinflussen. Die praktische Intelligenz haben sie in den Umgang mit Aufgaben, mit sich selbst und mit anderen aufgeteilt, wobei letzterer wiederum in den Umgang mit Vorgesetzten, mit Gleichgestellten und mit Unterstellten unterteilt ist. In Abbildung 2.5 ist ein Item aus dem *Tacit Knowledge Situational Judgment Test* (*TKSJT*) für Lehrpersonen dargestellt (tacit knowledge = implizites Wissen, Erfahrungswissen; Wagner, 1987).

Mr. Thompson usually gets along well with his colleagues. One day, in a departmental meeting about the curriculum, a colleague personally attacks him because Mr. Thompson expresses a different opinion about a new program than most of his colleagues.

Given the situation, please indicate in the box below what would be your primary concern in dealing with the situation.

Given the situation, please rate the quality of the following statements.

1	2	3	4	5	6	7
Strongly Disagree			Neutral			Strongly Agree

1. [COMPLY] Mr. Thompson should reiterate his opinion about the curriculum but state that he is willing to go along with the group.
2. [CONSULT] After the meeting, Mr. Thompson should ask one of the other teachers how he or she thinks he should deal with his colleague's comments.
3. [CONFER] Mr. Thompson should talk privately with his colleague and say that he felt the personal attack was inappropriate and out of line.
4. [AVOID] Mr. Thompson should ignore the attack and continue his discussion with another teacher.
5. [DELEGATE] Mr. Thompson should ask the principal speak to the colleague about his behavior.
6. [LEGISLATE] Mr. Thompson should propose the establishment of formal rules of order for faculty meetings.
7. [RETALIATE] Mr. Thompson should state that he is not interested in responding to petty personal attacks, but will be happy to answer questions about his opinion of the program.

Abbildung 2.5 Beispiel eines Items aus dem *Tacit Knowledge Situational Judgment Test* (*TKSJT*) für Lehrpersonen in High Schools (Stemler & Sternberg, 2006, S. 125).

Stemler und Sternberg (2006) benutzten wenn immer möglich die Angaben aus den Interviews, um die Verhaltensalternativen zu formulieren, wobei sie in der definitiven Version des TKSJT zu einzelnen Strategien auch zwei oder mehr Verhaltensweisen pro Item-Stamm aufführten. Zur Überprüfung dieses Testverfahrens führten sie eine Diskriminanzanalyse durch, da auf Grund der Abhängigkeit der Effektivitätseinschätzung einer Verhaltensweise von der geschilderten Situation die Durchführung einer explorative Faktorenanalyse ungeeignet ist (McDaniel & Nguyen, 2001).

CALIBRATOR ROLE:

Definition: Behaviors that function to observe the team social processes, to make the team aware of them, and to suggest changes to these processes that would bring them in line with functional social norms. The Calibrator role involves overt creation of new team norms dealing with team process issues (not task issues). It may involve initiating discussion of power struggles or tensions in the team, settling disputes among team members, summarizing team feeling, and soliciting feedback.

Conditions in which role is appropriate: Nonfunctional team processes: Represent situations in which functional patterns of social interaction have not been established in the team, or they have been disrupted by malfunctional behavior. Occurs when team is new and team members have little experience working together, or there are changes in team …; there is emotional or task-based conflict or distrust in the team …; work is "negotiation" oriented, and the context is socially demanding …

SJT-ITEM:

You are a member of a sales team at a local bookstore, where recent sales have been decreasing substantially due to a shrinking number of customers. You are in a team meeting discussing solutions to the declining sales problem. The discussion becomes a bit heated when the oldest team member suggests that the sales numbers for the new sales reps are quite low. One of the younger reps quickly counters that every time he asks for help with a customer, the older rep takes credit for the sale. The other new sales rep simply looks at the floor and says nothing. Please rate the effectiveness of each of the following responses. [Antwort-Skala: very ineffectiv (1), somewhat ineffective, neutral, somewhat effective, very effective (5), Anm. d. Verf.]

Get the quiet new sales rep involved by asking if she has noticed that the older sales rep has taken some of her sales as well. [ROLE INCONSISTENT]

Remind the two sales reps that personal attacks are not appropriate and that the team should focus on the future solutions.

Support the new team members by taking their side to make sure they are not used as "scapegoats" for the team's problems. [ROLE INCONSISTENT]

Remind the team that making critical remarks about specific people makes people defensive and will prevent the members from accomplishing anything as a team.

…

Abbildung 2.6 Definition der Calibrator-Rolle und Beispiel aus dem Team-Rollen-Test von Mumford et al. (2008, S. 255 resp. S. 267).

Ein ähnliches Vorgehen haben Mumford et al. (2008) für die Entwicklung ihres *Team Role Knowledge Situational Judgment Test* gewählt: Ausgehend von einer Typologie der Teamrollen (Mumford, Campion & Morgeson, 2006) formulierten sie zu jeder der zehn Rollen (Contractor, Creator, Contributor, Completer, Critic, Cooperator, Communicator, Calibrator, Consul und Coordinator) ein auf die betreffende Rolle abgestimmtes Szenario, wobei sie Teamsituationen aus unterschiedlichen Berufszweigen wählten. Zu jedem Szenario entwickelten sie daraufhin sechs bis zwölf unterschiedliche Verhaltensalternativen, wobei einige davon konsistent mit der zum Szenario passenden Rolle sind und so als richtige Antworten gelten (siehe Abbildung 2.6). Die vier Testscores (Aufgaben-Rollen, soziale Rollen, brückenschlagende Rollen und Gesamtwert) ergeben sich aus der Aufsummierung der rollenkonsistenten Einstufungen und der umcodierten Einstufungen aus den rolleninkonsistenten Verhaltensalternativen. Validitätsstudien ergaben eine minderungskorrigierte Korrelation von .45 zwischen dem Gesamtwert und der Einschätzung der Teamrollen-Leistung durch Peers. Die Korrelationen mit dem Gesamtwert und den Big Five-Dimensionen liegen zwischen -.10 und .16, diejenige mit Intelligenz beträgt .28.

Schritt 6: Wahl der Antwortinstruktionen

An der 14-ten Konferenz der Society for Industrial and Organizational Psychology 1999 führte Clevenger ein Symposium unter dem Titel „*The construct validity of the situational judgment inventory*" durch, um mit Experten über die Faktorenstruktur und die Korrelate von SJT zu diskutieren. Es zeigte sich, dass die fünf präsentierenden Forscherteams von unterschiedlich hohen Zusammenhängen der SJT zu kognitiver Leistungsfähigkeit und Persönlichkeit berichteten, was eine Synopse verunmöglichte. Als einen Grund dafür erkannten die Teilnehmer die unterschiedlichen Instruktionen, welche sie den Testnehmern bei der Bearbeitung der SJTs gegeben haben: Pereira und Schmidt Harvey (1999) zum Beispiel gaben die Anweisung, die Effektivität jeder Verhaltensalternative anzugeben, wohingegen bei Reynolds et al. (1999) die beste ausgewählt werden musste (Ployhart & Ehrhart, 2003). Dass Unterschiede bei den Instruktionen und der Verrechnung grosse Einflüsse auf die Konstruktvalidität haben können, zeigte sich auch schon bei biografischen Fragebogen und Interviews (z. B. Campion, Palmer & Campion, 1997). Diese Erkenntnisse führten dazu, dass Forschergruppen damit begannen, die Auswirkungen der Antwortinstruktionen auf die Validität von SJT empirisch zu untersuchen. McDaniel et al. (2006; siehe auch McDaniel, Hartmann, Whetzel & Grubb, 2007; Polyhart & Ehrhart, 2003) haben eine Taxonomie der Antwortinstruktionen von SJTs erstellt, die in Tabelle 2.1 abgebildet ist. Sie

unterscheiden zwischen Fragen nach der Verhaltenstendenz (Was würden Sie in dieser Situation am ehesten tun?) und nach dem Wissen (Was ist das beste Verhalten in dieser Situation?), was mit dem Konzept des typischen versus maximalen Verhaltens (Cronbach, 1949) vergleichbar ist.

Tabelle 2.1

Taxonomie der Antwortinstruktionen (nach McDaniel et al., 2006, 2007)

	Verhaltenstendenz (*behavioral tendency; would do*)	Wissen (*knowledge; should do*)
eine auswertbare Antwort	Wie würden Sie sich am ehesten verhalten?	Wählen Sie die beste / effektivste Antwort.
	Was haben Sie getan?	Wie sollte man sich hier verhalten?
zwei auswertbare Antworten	Wie würden Sie sich am wahrscheinlichsten resp. unwahrscheinlichsten verhalten?	Wählen Sie die beste und die schlechteste / zweitbeste Antwort.
so viele auswertbare Antworten wie Antwortoptionen	Stufen Sie bei jeder Antwort die Wahrscheinlichkeit ein, dass Sie sich so verhalten würden.	Stufen Sie die Effektivität jedes Verhaltens ein.
	Rangieren Sie die Antworten von der wahrscheinlichsten zur unwahrscheinlichsten.	Rangieren Sie die Verhaltensweisen von der besten zur schlechtesten.
	Stufen Sie die Wahrscheinlichkeit des Verhaltens auf einer Likertskala ein.	Wählen Sie die beste, zweit- und drittbeste Antwort.
		Stufen Sie die Wichtigkeit ein.

Ployhart und Ehrhart (2003) haben die Auswirkungen verschiedener, in früheren Studien verwendeten Instruktionen auf die Validität eines SJTs untersucht, indem sie denselben Test von knapp 500 Studenten mit folgenden sechs unterschiedlichen Instruktionen ausfüllen liessen:

A. Instruktionen zur Erfassung der Verhaltenstendenz (*would do*)

1. Wie haben Sie sich in der Vergangenheit typischerweise in einer solchen Situation verhalten? (z. B. Motowidlo et al., 1992)
2. Was würden Sie in dieser Situation am wahrscheinlichsten, was am unwahrscheinlichsten tun? (z. B. Motowidlo et al., 1990)
3. Stufen Sie die Wahrscheinlichkeit ein, mit welcher Sie die einzelnen Verhaltensweisen ausführen würden. (z. B. Bruce & Learner, 1958; Hunter, 2003)

B. Instruktionen zur Erfassung des Wissen (*should do*)

1. Was sollte in dieser Situation getan werden? (z. B. Phillips, 1992, 1993)
2. Welche Verhaltensweise beurteilen Sie als die effektivste, welche als die ineffektivste? (z. B. Weekley & Jones, 1999)
3. Stufen Sie die Effektivität jeder Verhaltensweise ein. (z. B. Jones, Dwight & Nouryan, 1999)

Die *should do*-Versionen ergaben höhere Skalenmittelwerte, geringere Standardabweichungen und führten zu Verteilungen, welche mehr von der Normalverteilung abweichen als bei den *would do*-Instruktionen. Die interne Konsistenz (Cronbach Alpha) ist bei den Einstufungs-Items am höchsten (Werte von .65 bis .73), gefolgt von den Items, bei welchen zwei Antworten verlangt werden (Werte von .30 bis .65). Tendenziell die tiefsten Reliabilitäten zeigten Skalen, bei denen nur eine Antwort gegeben werden musste (.24 bis .65). „*Would do*"-SJT ergaben zudem höhere Retest-Reliabilitäten als „*should do*", wobei der höchste Wert dasjenige Format erzielte, bei welchem die Testbearbeiter bei jeder Alternative die Verhaltenswahrscheinlichkeit ankreuzen mussten (r_{tt} = .92). Auch bei den Kriteriumsvaliditäten zeigten sich Unterschiede zwischen den beiden Formaten: *Would do*-Skalen korrelierten deutlich höher mit dem Notendurchschnitt im College (r = .43 bis .52 vs. r = .07 bis .22) und mit der Selbst- und Fremdeinschätzung der Leistung (r =.64 bis .89 vs. r = -.02 bis .29) als *should do*-Skalen. Auf Grund ihrer Resultate geben Ployhart und Ehrhart (2003) den Rat, dass in einem SJT nur ein einziges Antwortformat zu verwenden ist, da der Einsatz unterschiedlicher Formate zu Mehrdimensionalität führen und somit die Konstruktvalidität beeinflussen und die Kriteriumsvalidität verringern könnte.

Andere Resultate zeigten sich in der Studie von Nguyen und McDaniel (2003; siehe auch McDaniel & Nguyen, 2001): Bei der Wissensinstruktion (*should do*) fiel die Korrelation mit Intelligenz höher aus, als bei der Verhaltenstendenz-Anweisung (r = .39 vs. r = .23). Dieser Befund zeigt sich auch in der Meta-Analyse von McDaniel et al. (2007, ρ = .35 resp. ρ = .19). Bei Persönlichkeitsdimensionen ergibt sich jedoch – ausser bei Extraversion (ρ = .15 vs. ρ = .08) – ein umgekehrter Zusammenhang: Hier korrelieren die Skalen mit der *should do*-Anweisung höher als diejenigen mit der *would do*-Anweisung (Verträglichkeit: ρ =.19 vs. ρ =.37; Gewissenhaftigkeit: ρ =.24 vs. ρ =.34; emotionale Stabilität: ρ =.12 vs. ρ =.35). Kein Unterschied zeigte sich bei der Offenheit für Erfahrung und bei der Arbeitsleistung im Beruf. *Would do*-Anweisungen ergeben zudem höhere Korrelationen mit kognitiver Leistungsfähigkeit als mit Persönlichkeit (McDaniel et al., 2007). Für McDaniel et al. stellen SJTs somit die

einzige Verfahrenskategorie dar, bei welcher je nach Instruktion typisches (*would do*-Anweisungen) oder maximales Verhalten (*should do*-Anweisungen) gemessen werden kann.

SJT mit Wissensinstruktionen (*should do*) sind resistenter gegenüber bewusster Antwortverfälschung (*faking*), da Bewerber merken, dass das, was sie üblicherweise tun würden, nicht die effektivste Verhaltensweise ist und deshalb eine andere Lösung wählen (McDaniel & Nguyen, 2001; Nguyen, Biderman & McDaniel, 2005). Nach Paulhus (1984) sind maximale und typische Prädiktoren unterschiedlich anfällig auf Verzerrungen, welche mit Selbsttäuschung oder Impression Management zusammenhängen. Diese treten eher bei typischen Verhaltensbeschreibungen (*would do*) auf.

Schritt 7: Entwicklung des Auswertungsschlüssels (keying and scoring)

Eine weit verbreitete, schon von Motowidlo et al. (1990) eingesetzte Methode zur Festlegung des Auswertungsschlüssels für SJTs besteht darin, dass mindestens fünf bis sieben Subject Matter Experts – Motowidlo et al. befragten mehr als 30 – mit langjähriger Berufserfahrung im betreffenden Aufgabengebiet die Effektivität jeder Verhaltensalternative auf einer Skala einstufen (McDaniel & Whetzel, 2007). Dabei sind Items respektive Verhaltensalternativen auszuschliessen, bei welchen sich die Experten hinsichtlich der Effektivität uneinig sind (Motowidlo et al., 1997). In der Fachliteratur werden zusätzlich eine Vielzahl anderer Auswertungsstrategien beschrieben, was Bergman, Drasgow, Donovan, Henning und Juraska (2006) treffend im Titel ihres Artikels ausdrücken: „Scoring situational judgment tests: Once you get the data, your troubles begin." Dies, da sich bei der Auswertung von SJT im Gegensatz zu den meisten anderen für die Personalselektion eingesetzten Verfahren die Frage nach dem optimalen Vorgehen stellt, da SJT-Items häufig keine objektiv richtige Antwort haben und mehrere, wenn nicht alle Antwortalternativen plausibel, jedoch unterschiedlich effektiv sind (Bergman et al., 2006; Hanson, Borman, Mogilka, Manning & Hedge, 1999). Somit stellt sich die Frage nach der besten, nicht nach der richtigen Antwort. Eine ähnliche Problematik ergibt sich auch bei der Auswertung von biografischen Fragebogen, wodurch man sich bei der Festlegung der Auswertungsstrategie bei SJTs auf die in diesem Gebiet erzielten Forschungsergebnisse abstützen kann (z. B. Devlin, Abrahams & Edwards, 1992; Hogan, 1994). Dass die Wahl des richtigen respektive angemessenen Auswertungsschlüssels nicht trivial ist, zeigt sich darin, dass die unterschiedlichen Arten des Scorings die Reliabilität und die Validität (z. B. Krokos, Meade, Cantwell, Pond & Wilson, 2004) beeinflussen. Als Beispiel stelle ich in Tabelle 2.2 die Ergebnisse der Studie von Hanson und Borman

(1995) vor, welche die Reliabilität von fünf verschiedenen Scoringmethoden eines 35 Items umfassenden SJT zu militärischem Führungsverhalten untersuchten.

Tabelle 2.2

Reliabilitäten (Cronbach Alpha) einer Skala zu militärischem Führungsverhalten beim Einsatz unterschiedlicher Scoring-Methoden (nach Hanson & Borman, 1995)

Scoring-Methode	α
Anzahl richtig eingeschätzte effektivste Verhaltensalternativen	.60
Verrechnung der anhand des Expertenratings bestimmten Effektivitätsgewichtung der als am effektivsten gewählten Antwort	.68
Anzahl richtig eingeschätzte ineffektivste Verhaltensalternativen	.57
Verrechnung der anhand des Expertenratings bestimmten Effektivitätsgewichtung der als am ineffektivsten gewählten Antwort	.68
Subtraktion des Effektivitätsscores der als ineffektivesten von der als effektivsten gewählten Verhaltensalternative	.75

Bei der Anwendung des „klassischen" Auswertungsverfahrens für SJTs erhält der Testbearbeiter einen Punkt, wenn seine Antwort mit der Experteneinschätzung übereinstimmt, bei einer entgegengesetzten Antwort einen Minuspunkt, in allen anderen Fällen null Punkte (McDaniel et al., 2006; Motowidlo et al., 1990; Ployhart & Ehrhart, 2003; Weekley & Jones, 1999). Somit reicht die Streubreite der möglichen Punktzahlen bei einer „Was soll getan werden?"-Anweisung von –1 bis +1 und bei einer „Wähle die beste und die schlechteste Antwort"-Anweisung von –2 bis +2, was den Vorteil grösserer Varianz bringt (Weekley, Harding, Creglow & Ployhart, 2004). In mehreren Studien zeigte sich jedoch ein psychometrischer Unterschied zwischen der Best- und der Schlechtest-Antworten: Cucina, Vasilopoulos und Leaman (2003) bildeten je einen Score für die *most likely*-Antwort und die *least likely*-Antwort, womit sie eine bessere Validität erzielten. Zudem konnten sie aufzeigen, dass die Best-Wahl höher mit den Big Five korreliert als die Schlechtest-Wahl. Zu einem vergleichbaren Ergebnis kamen McElreath und Vasilopoulos (2002), bei welchen die least likely-Antworten höher mit Intelligenz korrelierten als die most likely-Antworten.

Bergman et al. (2006; siehe auch Weekley et al., 2006) unterteilen die in der Literatur beschriebenen Scoring-Methoden in die drei Kategorien empirisches, theoretisches und expertenbasiertes Scoring und führen noch eine vierte Kategorie ein, das Hybrid-Scoring.

Beim *empirischen Scoring* – einem bei biografischen Fragebogen üblichen Vorgehen (z. B. Hogan, 1994) – werden die Einstufungen der verschiedenen Verhaltensalternativen anhand ihres Zusammenhangs mit einem Kriterium gewichtet (Mumford & Owens, 1987; Mumford & Whetzel, 1997). So können zum Beispiel die Einstufungen der einzelnen Verhaltensweisen pro Item mit der Arbeitsleistung korreliert und diejenige mit der höchsten Validität als „richtige" Antwort gewählt werden. Bei diesem Vorgehen ersetzen also die leistungsstarken Jobinhaber die beim expertenbasierten Scoring eingesetzten SMEs, was dazu führt, dass Verhaltensalternativen, welche gut zwischen leistungsstarken und leistungsschwachen Jobinhabern unterscheiden, höhere Gewichte erhalten, als andere Alternativen, auch wenn diese als bessere Antworten angesehen werden könnten. Und so werden auch offensichtliche, transparente Antworten, welche sowohl die leistungsstarken wie auch die leistungsschwachen Testbearbeiter wählen, nicht gewichtet, weil sie nicht zwischen diesen beiden Gruppen unterscheiden können. Voraussetzung für einen erfolgreichen Einsatz dieser Methode ist natürlich, dass das gewählte Kriterium die zu messende Leistungsfähigkeit auch vollständig abdeckt. Aus diesem Grunde müssen die Testentwickler beim Einsatz des empirischen Scorings immer die Reliabilität und Validität des Kriteriums überprüfen (Mumford, 1999). Steht ein diesen Ansprüchen genügendes Kriterium nicht zur Verfügung, stellen SMEs die bessere Wahl dar.

Empirisches Scoring bei der Entwicklung eines SJTs setzen zum Beispiel Dalessio (1994) oder Lievens (2000) ein. Auch Weekley und Jones (1997) haben sich für ein empirisches Scoring bei ihrem videobasierten SJT für die Selektion von Verkaufspersonal entschieden. Zum anhand von 15 Fokusgruppen-Interviews erstellten Anforderungsprofil formulierten sie 50 Situationen mit je vier Verhaltensalternativen. Um das empirische Scoring zu bestimmen, liessen sie knapp 700 Angestellte den SJT bearbeiten und zusätzlich deren Arbeitsleistung von deren direkten Vorgesetzten einstufen. Dazu setzten sie eine 47 Items umfassende Tätigkeitsliste mit einer vierstufigen Likert-Skala (4 = exzellent bis 1 = schlecht) ein. Pro Verhaltensalternative berechneten Weekley und Jones den Durchschnitt der fremdbeurteilten Arbeitsleistung derjenigen Angestellten, welche diese Alternative gewählt haben. Die Verhaltensalternative mit der höchsten durchschnittlichen Arbeitsleistung codierten sie daraufhin mit 1, diejenige mit dem tiefsten Arbeitsleistung mit –1 und die beiden restlichen mit null. Sieben Items mussten sie ausschliessen, weil sie diese nicht mit dem gewählten Schema codieren konnten.

Mit einem Verweis auf Forschungsergebnisse zu klinischer und statistischer Prädiktion (z. B. Grove, Zald, Lebow, Snitz & Nelson, 2000; Kleinmuntz, 1990; Meehl, 1954) spricht sich Lievens (2000) für die Verwendung von empirischem

Scoring aus und schlägt zur Auswertung der Daten die Korrespondenzanalyse (z. B. Greenacre, 2007; Kiers, 1991) vor. Ein grosser Vorteil dieses Vorgehens sei zudem die Verwendung von Gewichten bei der Berechnung der Gesamtscores. Beim Leadership-SJT von Krokos et al. (2004) zeigte sich zudem, dass der Einsatz von empirischen Methoden bei der Bestimmung des Auswertungsschlüssels (*vertical & horizontal percent methods*, *correlational method, mean criterion method*) deutlich höhere Validitätskoeffizienten erbringt, als die SME-Methode (r = .05 - .28 vs. r = -.15).

SJTs können auch vor dem Hintergrund einer Theorie konstruiert werden und lassen sich somit anhand eines *theoretischen* respektive *konstruktbasierten Scorings* auswerten. Die Entwicklung einer konstruktbasierten SJT-Skala verläuft dabei gleich, wie diejenige einer herkömmlichen Persönlichkeitsskala, indem der Testentwickler zur zu messenden Dimension Items konstruiert und anhand einer Skalenanalyse Items ausschliesst, welche nur schwach mit der Gesamtskala korrelieren (Mumford, 1999). Die Auswertung eines solchen SJTs weist jedoch Besonderheiten auf: So ergibt eine vom Testbearbeiter gewählte Verhaltensalternative, welche der Theorie entspricht einen Punkt, eine solche, welche der Theorie widerspricht einen Minuspunkt. Bergman et al. (2006) führen dazu ein Beispiel zum Delegationsstil auf: Sie unterscheiden zwischen Empowerment (die Gruppe fällt die Entscheidung), Partizipation (die Führungskraft delegiert die Entscheidung ans Team) und Aufgabenorientiertheit (die Führungskraft entscheidet). Im SJT haben sie diese drei Delegationsstile jeweils mit einem Score abgebildet, wobei sie zum Beispiel beim Empowerment-Score Antworten, welche dieses Verhalten beschreiben als richtig werteten, Antworten zur Aufgabenorientiertheit als falsch und alle anderen Antworten als irrelevant. Gemäss Hough und Paullin (1994) sind theoretische Scorings auf Grund ihrer Transparenz anfällig für Faking.

Beim *experten-basierten Scoring* erstellen die Testentwickler den Auswertungsschlüssel auf Grund der Einstufungen von Experten, das heisst Personen mit langjähriger Berufserfahrung (McDaniel & Nguyen, 2001). Bei Forced-Choice-Verfahren erhält dabei der Testbearbeiter für eine richtige – das heisst für eine mit dem Expertenurteil übereinstimmende – Antwort ein Punkt zugeordnet (z. B. Hunter, 2003; Stevens & Campion, 1999). Da die Auswahl einer Antwort bei Forced-Choice-Verfahren zu kategorialen Daten führt, welche sich nicht mehr mit den statistischen Standardmethoden bearbeiten lassen (Lievens, 2000), setzen viele Testentwickler auch Likert-Skalen ein (Weekley et al., 2006) und verwenden zum Teil komplexe Auswertungsmethoden. McDaniel et al. (2006) schlagen den Einsatz einer vierstufigen Rating-Skala zur Einstufung der Effektivität jeder Antwortalternative vor. Sie sehen darin den Vorteil, dass den Bewerbern ein

grösseres Antwortspektrum angeboten wird und man sie nicht zu einer dichotomen Antwort zwingt. Übereinstimmungen mit dem Expertenurteil scoren sie mit +1, Abweichungen mit –1. Hingegen raten sie von einer Experteneinstufung auf einer vierstufigen Skala ab, da diese von einer Unterscheidung zwischen effektiv und sehr effektiv häufig überfordert sind.

Auch Krokos et al. (2004) sprechen sich dagegen aus, bei einem SJT „richtige" und „falsche" Antworten zu bestimmen. Dies sei nur bei SJTs angebracht, welche Intelligenz oder einfaches Wissen erfassen. Die meisten SJTs beziehen sich jedoch auf komplexe, soziale oder praktische Aspekte von Leistung in Arbeitssituationen. Sie gehen davon aus, dass sich die verschiedenen Verhaltensalternativen auf einem Kontinuum von effektiv bis ineffektiv anordnen lassen. Vor allem bei komplexeren Tätigkeiten, wie bei Managern oder Fachspezialisten, gibt es häufig nicht richtiges oder falsches Verhalten, sondern nur mehr oder weniger effektives. Dem wird zum Teil Rechnung getragen, indem die Testentwickler SMEs die Effektivität der jeweiligen Antwortalternativen beurteilen lassen und die Interraterreliabilität berechnen (Hanson et al., 1999). Items mit Alternativen mit tiefer Interrater-Übereinstimung werden überarbeitet oder ausgeschlossen. So liessen zum Beispiel Weekley et al. (2004) die Antworten rangieren und verglichen die Rangfolge der Testbearbeiter mit denjenigen der SMEs mittels Korrelation.

Hanson et al. (1999) tragen beim Scoring in ihrem SJT dem Umstand Rechnung, dass nicht alle ineffektiven Antworten gleich schlecht sind und liessen die Effektivität der einzelnen Verhaltensalternativen von SMEs einschätzen, um so Gewichte für die Antworten bestimmen zu können. Der Testbearbeiter erhält so nicht einfach einen Punkt für die richtige Antwort und keinen Punkt für eine falsche, sondern den jeweiligen Effektivitätslevel der gewählten Antwort, was zu einer höheren Reliabilität führt als das herkömmliche Rating (Hanson & Borman, 1995; siehe auch Tabelle 2.2). Ein ähnliches Verfahren wählten Legree, Psotka, Tremble und Bourne (2005), indem sie den Mittelwert der Effektivitätseinschätzung pro Verhaltensalternative berechnen und entsprechend der Höhe der Abweichung der Antwort von diesem Mittelwert Minuspunkte geben. So erhält ein Testbearbeiter zum Beispiel für die Wahl des Effektivitätslevels 4 bei einer von den SMEs durchschnittlich mit einer Effektivität von 3.25 beurteilten Verhaltensalternative -0.75 Punkte (*data-assisted rational keying*). Chan und Schmitt (2002) berechnen bei ihren sechsstufigen Likert-Skalen die Übereinstimmung der Effektivitätseinschätzung in Prozent und vergeben für jede Aufgabe je nach Grad der Übereinstimmung 0 bis 2 Punkte.

McDaniel et al. (2006) machen darauf aufmerksam, dass die SMEs bei der Einstufung der Verhaltensalternativen unter Umständen auf Firmennormen oder firmenspezifische Vorgaben referenzieren, welche Bewerbern nicht bekannt sein dürften. Somit muss bei diesem Vorgehen kontrolliert werden, ob sich bei den einzelnen Items grosse Differenzen bei der Einschätzungen durch die Stelleninhaber und die Bewerber ergeben. Ist dies der Fall, so ist das entsprechende Item zu entfernen oder umzuformulieren.

Ein Nachteil bei der Expertenbeurteilung ist, dass es häufig vorkommt, dass sich die SMEs nicht darüber einig sind, welches die beste Handlungsalternative darstellt (Lievens, 2000). Dies führt dazu, dass auf Grund der benötigten Übereinstimmung zwischen den Experten die Gefahr besteht, dass die richtigen Antworten auch die transparentesten sein werden. Zudem hängt die Qualität und die Aussagekraft stark mit der Auswahl der SMEs und deren Vorstellung vom zu messenden Konstrukt zusammen. Tiefe Validitätswerte von SME-SJTs können neben der Möglichkeit, dass die SME-Ratings unzutreffend sind, auch damit zusammenhängen, dass sich die Vorstellung vom zu messenden Konstrukt zwischen den Experten und den Bewerbern unterscheidet (Krokos et al., 2004). So ist es gut möglich, dass einige der Bewerber gewisse der in einem SJT geschilderten Situationen aus dem Berufsalltag noch nie erlebt haben.

Es stellt sich nun die Frage, mit welcher der drei oben referierten Scoring-Methoden sich die besten Ergebnisse erzielen lassen. Mumford (1999) schreibt dazu:

> Each of these scaling techniques reflects a different approach, embodying different assumptions, with different objectives. Nonetheless, these techniques are not necessarily mutually exclusive, and it may well be the case that a judicious combination of scaling strategies will result in stronger background data measures. (S. 129)

Damit spricht Mumford das *Hybrid-Scoring* an, welches dann vorliegt, wenn die Testentwickler verschiedene Scoring-Methoden kombinieren. Mumford (1999) und Olson-Buchanan et al. (1998) gehen davon aus, dass sich mit einem Hybrid-Scoring die prädiktive Validität eines SJTs verbessern lässt. Bergman et al. (2006) geben als Beispiel für ein Hybrid-Scoring an, dass bei der Verwendung von zwei Auswertungsschlüsseln ein positiver Wert in einem einen negativen Wert im anderen kompensieren kann. Im Prinzip können alle oben dargestellten Scoring-Methoden kombiniert werden. Es müssen dabei jedoch theoretische (z. B. „Macht es auf Grund von theoretischen Überlegungen Sinn, diese beiden Auswertungsschlüssel zu kombinieren?“) und praktische (z. B. „Können die beiden Auswertungsschlüssel kombiniert werden, ohne dass eine zusätzliche Validie-

rungsstudie durchzuführen ist?") Fragen beachtet werden (Bergman et al., 2006).

Abschliessend zu dieser Darstellung der Scoring-Methoden zitiere ich noch Bergman et al. (2006), welche auf die Wichtigkeit der Überprüfung verschiedener Scoring-Methoden bei der Entwicklung eines SJTs hinweisen:

> Keys should be assessed for validity, incremental validity, adverse impact, and construct validity ... From these analyses, the best key(s) can be identified. Although this validation strategy seems basic, studies in the SJT literature have rarely addressed the potential differential validity of the multiple keys available for a given test. As demonstrated here, it is essential that researchers critically evaluate their SJT keying choices. (S. 233)

2.3 Zusammenhänge von Situational Judgment Tests mit Arbeitsleistung, Intelligenz und Persönlichkeit

Bislang noch nicht abschliessend geklärt ist die Frage, welche Konstrukte SJTs messen. Da die Scores aus SJTs mit Arbeitsleistung, Intelligenz und Persönlichkeitseigenschaften korrelieren, gelangten Smith und McDaniel (1998) zum Schluss, dass SJTs verschiedene arbeitsbezogene Konstrukte erfassen. Dabei weisen SJTs inkrementelle Validität zu Intelligenz, Berufswissen, Berufserfahrung, Gewissenhaftigkeit und weiteren Persönlichkeitsdimensionen auf (z. B. Clevenger et al., 2001), wobei diese mit .03 bis .07 jedoch eher gering ausfällt (McDaniel et al., 2007). In Tabelle 2.3 habe ich die Ergebnisse der Regressionsanalyse von Chan und Schmitt (2002) dargestellt, welche die inkrementelle Validität von SJTs übersichtlich aufzeigt. Die Varianzaufklärung nimmt unter Einbezug von SJT signifikant zu und zwar sowohl bei der Arbeitsleistung (*overall job performance*) als auch bei deren Subdimensionen (Berufskenntnisse, Engagement im Beruf und soziale Kompetenz). Die Studie von Chan und Schmitt (2002) unterstützt dieses Ergebnis, indem sie aufzeigt, dass SJTs ein stabiles Konstrukt messen, das nur schwach (r = -.10) mit Berufserfahrung und nicht mit Intelligenz (r = -.02) zusammenhängt und sich von den Big Five-Persönlichkeitseigenschaften unterscheidet ($|r|$ = .19 bis .29). Smith und McDaniel (1998) fanden die höchsten Korrelationen ihres SJTs mit Alter und Berufserfahrung, was dafür spricht, dass ihr Verfahren arbeitsplatzbezogenes Wissen erhebt, welches durch Arbeits- und Lebenserfahrung gewonnen wird.

Da die Korrelationen von SJTs mit kognitiver Leistungsfähigkeit und Persönlichkeitsaspekten von Studie zu Studie variieren (McDaniel & Nguyen, 2001), ist anzunehmen, dass sie von einem oder mehreren Moderatoren beeinflusst werden. McDaniel et al. (2007) schlagen auf Grund ihrer Ergebnisse vor, dass bei der Personalselektion immer ein Intelligenztest eingesetzt werden sollte, da dieser nachweislich einer der besten Prädiktoren für Berufserfolg ist. Möchte der Personalverantwortliche einen zusätzlichen Test einsetzen, so kann er zwischen einem Big Five-Persönlichkeits-Fragebogen und einem SJT wählen, da beide ungefähr dasselbe Ausmass an inkrementeller Validität leisten.

Tabelle 2.3

Hierarchische Regressionsanalyse von Intelligenz, Persönlichkeit, Berufserfahrung und SJT auf Leistungskriterien (nach Chan & Schmitt, 2002, S. 247)

Schritt	Prädiktor	Berufs-kenntnisse	Engagement im Beruf	Soziale Kompetenz	Arbeits-leistung
1	Intelligenz	.24*	.03	-.01	-.03
	Gewissenhaftigkeit	.09	.24*	-.05	.15
	Extraversion	.06	-.03	.26*	-.01
	Verträglichkeit	.03	.00	.01	.12
	Neurotizismus	.11	.04	-.10	.15
	Offenheit für Erfahrung	.21*	.17*	.21*	.17*
	Berufserfahrung	.01	-.07	.05	-.10
	R^2	**.15***	**.13***	**.20***	**.11***
2	Situational Judgment Test	.24*	.30*	.16*	.22*
	R^2	**.20***	**.21***	**.23***	**.15***
	delta R^2	**.05***	**.08***	**.03***	**.04***

Anmerkung. N = 160. In der Tabelle sind die Beta-Werte aufgeführt. * $p < .05$.

Eine von McDaniel et al. (2001) durchgeführte Meta-Analyse, welche 79 Studien mit Ergebnissen zum Zusammenhang zwischen SJTs und Intelligenz umfasst, ergab eine durchschnittliche Korrelation von ρ = .46. Es zeigte sich zudem, dass SJTs, welche auf einer Arbeitsplatzanalyse basierten, höher mit Intelligenz korrelieren (ρ = .50) als solche ohne entsprechende Analyse (ρ = .38). Zudem konnten McDaniel, Hartman und Grubb (2003) in ihrer Meta-Analyse nachweisen, dass die Höhe der Korrelation auch mit der Test-Instruktion zusammenhängt: So betrug der durchschnittliche Zusammenhang bei einer Wissensinstruktion (*should do*) ρ = .35 (k = 69) und bei einer Verhaltensinstruktion (*would do*) ρ = .19 (k = 26). McDaniel et al. (2003) konnten in ihrer Studie auch aufzeigen, dass SJTs bezüglich Berufsleistung inkrementelle Validität gegenüber Intelligenztests

aufweisen. Beim Zusammenhang mit der Berufsleistung (ρ = .26, k = 118) zeigten sich jedoch keine Unterschiede bezüglich der Art der Instruktion (McDaniel et al., 2007).

Chan und Schmitt (2002) erklären die gute prädiktive Validität von SJTs und die inkrementelle Validität gegenüber anderen Messverfahren mit deren Multidimensionalität. Da die Berufsleistung auch ein sehr heterogenes Konstrukt ist, welches sich aus unterschiedlichstem Verhalten in verschiedensten Situationen definiert, ergibt sich eine grosse Überschneidung zu den in einem SJT erfassten Konstrukten. „Because most job performance situations are complex, good judgment in these situations is likely to be a function of multiple, more narrowly defined traits and abilities." (Chan & Schmitt, 2002, S. 233) Als weitere Erklärung führen sie die Konstruktionsmethode an, bei welcher die anhand einer Arbeitsanalyse bestimmten erfolgsrelevanten Situationen aus dem Arbeitsalltag in Item-Stämmen abgebildet sind. Somit ergibt sich der Zusammenhang zwischen einem SJT und der Berufsleistung beinahe zwangsläufig. Da man sich jedoch bei der Entwicklung eines SJTs häufig auf Erfahrungen von Jobinhabern abstützt, welche ein Bewerber nicht unbedingt teilt, ist davon auszugehen, dass sich die konkurrente und die prädiktive Validität eines SJTs systematisch unterscheiden (Hanson et al., 1998).

In Tabelle 2.4 habe ich die Korrelationskoeffizienten der beiden Meta-Analysen zum Zusammenhang zwischen SJTs und den Big Five von McDaniel et al. (2003, 2007) dargestellt. Von der Annhame ausgehend, dass SJT Berufsleistung messen, verwundert nicht, dass die Korrelationen mit emotionaler Stabilität, Verträglichkeit und Gewissenhaftigkeit am höchsten ausfallen, da es sich gezeigt hat, dass diese Persönlichkeitsdimensionen valide Prädiktoren von Berufsleistung sind (Barrick & Mount, 1991). Motowidlo, Borman und Schmitt (1997) erklären sich die Zusammenhänge zwischen SJTs und Persönlichkeitsaspekten anhand der Überlegung, dass die Einschätzung der Effektivität eines Verhaltens in einer bestimmten Arbeitssituation von verschiedenen, individuell unterschiedlich ausgeprägten Fähigkeiten wie Intelligenz, Berufswisssen und Persönlichkeitsfaktoren abhängt.

Eine andere Erklärung für den Zusammenhang zwischen Persönlichkeitseigenschaften und Berufsleistung liefert die *implicit trait policy* (ITP; Motowidlo et al., 2006a, 2006b): Gemäss der Neigungs-Passung (*dispositional fit argument*) entwickeln Menschen Überzeugungen zur Effektivität unterschiedlichen Verhaltens in Übereinstimmung mit den Grundzügen ihrer Persönlichkeit. Wenn nun die Überwindung einer schwierigen Situation bei der Arbeit den Ausdruck einer spezifischen Persönlichkeitseigenschaft erfordert, werden Personen, bei welchen diese

Eigenschaft hoch ausgeprägt ist, eher glauben, dass ihr Verhalten in dieser Situation erfolgsversprechend ist. Somit haben sie genaueres und korrekteres Wissen darüber, wie sie sich in dieser Situation effektiv zu verhalten haben und es wird angenommen, dass dieses Wissen den Effekt der relevanten Persönlichkeitseigenschaft auf die Leistung in dieser Situation mediiert. „The dispositional fit argument implies that when people judge the effectiveness of behavioral episodes that express varying levels of some personality trait, their standing on that trait interacts with trait levels expressed by the behaviors to affect their effectiveness judgments." (Motowidlo et al., 2006b, S. 751)

Tabelle 2.4
Meta-Analysen zum Zusammenhang von SJTs mit den Big Five

	McDaniel et al. (2003)		McDaniel et al. (2007)	
	Anzahl Studien	ρ	*Anzahl Studien*	ρ
Verträglichkeit	16	.33	51	.25
Wissensinstruktion	5	.20	34	.19
Verhaltensinstruktion	11	.53	17	.37
Gewissenhaftigkeit	19	.37	53	.27
Wissensinstruktion	8	.33	38	.24
Verhaltensinstruktion	11	.51	15	.34
Emotionale Stabilität	14	.41	49	.22
Wissensinstruktion	4	.11	33	.12
Verhaltensinstruktion	10	.51	16	.35
Extaversion	10	.20	25	.14
Wissensinstruktion	5	.21	14	.15
Verhaltensinstruktion	5	.11	11	.08
Offenheit für Erfahrung	5	.12	19	.13
Wissensinstruktion	1	.25	11	.14
Verhaltensinstruktion	4	.09	8	.11

Wird ein SJT nach der üblichen Vorgehensweise entwickelt, können jedoch keine hohen Korrelationen mit Persönlichkeitseigenschaften oder faktorenanalytisch interpretierbare Faktoren erwartet werden (McDaniel & Whetzel, 2005). Dies ist einerseits auf den Konstruktionsprozess zurückzuführen, welcher nicht darauf ausgerichtet ist, voneinander unabhängige, reliable Skalen zu generieren. Andererseits setzten die Testentwickler zur Generierung der Item-Stämme Nichtpsychologen ein, welche nicht über Wissen darüber verfügen, wie das von ihnen

beobachtete Verhalten kategorisiert und in reliablen Skalen abgebildet werden könnte. Diesen Schluss zogen Lievens und Conway (2001) anhand ihrer Meta-Analyse von 34 Studien zur Konstruktvalidität von Assessment Centern, in welcher sich zeigte, dass rigoros konstruierte Assessment Centers eher die gewünschten Dimensionsfaktoren abbildeten. Auch Trippe (2002) konnte nachweisen, dass es mit einer konsequent darauf ausgerichteten Testkonstruktion möglich ist, im SJT die intendierten Dimensionen valide abzubilden. McDaniel et al. (2007) regen dazu an, den Zusammenhang zwischen dem Inhalt eines SJTs und der Konstruktvalidität besser zu erforschen:

> We also encourage research on the specification of content assessed in SJTs and the relation between content and validity. It is reasonable to expect that some content will yield different criterion-related and construct validity than other content. We encourage research in test development technology so that SJTs can be written to achieve prespecified correlations with other measures. (S. 84)

Von der (bestätigten) Annahme ausgehend, dass SJT Arbeitsverhalten vorhersagen können, indem sie arbeitsbezogenes Wissen erfassen, ergeben sich zwei Voraussetzungen, welche erfüllt sein müssen, dass SJT als valide Prädiktoren eingesetzt werden können (Hanson et al., 1998): Erstens müssen die Situationen im SJT mit denen im Job vergleichbar sein, was bedeutet, dass ein SJT nur für einen bestimmten Job entwickelt werden kann. Zweitens muss der Testbearbeiter Erfahrungen in den im SJT geschilderten oder mit anderen vergleichbaren Situationen gesammelt haben, um überhaupt über das notwendige Wissen zu verfügen. Damit lässt sich auch eines der Ergebnisse der Meta-Analyse von McDaniel et al. (1997) erklären, dass weniger detailliert beschriebene Situationen eine höhere Validität aufweisen als Items bei welchen die Situation ausführlich beschrieben ist. Bei weniger spezifisch formulierten Items, werden mehr vergleichbare Situationen angedeutet, was die Wahrscheinlichkeit erhöht, dass der Testbearbeiter die geschilderte Situation so oder ähnlich schon erlebt hat und somit genauer Auskunft über sein Verhalten geben kann.

2.4 Literaturverzeichnis

Anderson, L., & Wilson, S. (1997). Critical incident technique. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measure methods in industrial psychology* (pp. 89–114). Palo Alto, CA: Davis Black.

Banki, S., & Latham, G. P. (2010). The criterion-related validities and perceived fairness of the situtional interview and the situational judgment test in an Iranian organisation. *Applied Psychology: An International Review, 59*, 124–142.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1–26.

Bauer, T. N., & Truxillo, D. M. (2006). Applicant reactions to situational judgment tests: Research and related practical issues. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 233–249). Mahwah, NJ: Erlbaum.

Becker, T. E. (1998). Integrity in organizations: Beyond honesty and conscientiousness. *Academy of Management Review, 23,* 154–161.

Becker, T. E. (2000, August). *Hallmarks and consequences of integrity in organizations: The employees' perspective.* Paper presented at the 2000 annual Academy of Management Meetings, Toronto.

Becker, T. E. (2005). Development and validation of a situational judgment test of employee integrity*. International Journal of Selection and Assessment, 13*, 225–232.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., & Juraska, S. E. (2006). Scoring situational judgment tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment, 14*, 223–235.

Bledow, R., & Frese, M. (2009). A situational judgment test of personal initiative and its relationship to performance. *Personnel Psychology, 62*, 229–258.

Bruce, M. M. (1965). *Examiner's manual: Business judgment test*. Larchmont, NY: Author.

Bruce, M. M., & Learner, D. B. (1958). A supervisory practices test. *Personnel Psychology*, *11*, 207–216.

Butterfield, L. D., Borgen, W. A., Amundson, N. E., & Maglio, A. T. (2005). Fifty

years of the critical incident technique: 1954–2004 and beyond. *Qualitative Research, 5*, 475–497.

Campion, M. A., Palmer, D. K., & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology, 50*, 655–702.

Cardall, A. J. (1942). *Preliminary manual for the test of practical judgment*. Chicago: Science Research Associates.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82,* 143–159.

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15,* 233–254.

Chan, D., & Schmitt, N. (2005). Situational judgment tests. In A. Evers, N. Anderson & O. Voskuijl (Eds.), *The Blackwell handbook of personnel selection* (pp. 219–242). Malden, MA: Blackwell.

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82*, 300–310.

Christian, M. S., Edwards, B. D., & Bradley, J. C. (2010). Situational judgment tests: Constructs assessed and a meta-analysis of their criterion-related validities. *Personnel Psychology, 63*, 83–117.

Clevenger, J. P., Jockin, T., Morris, S., & Anselmi, T. (1999, April). *A situational judgment test for engineers: Construct and criterion related validity of a less adverse alternative.* Paper presented at the 14th Annual Conference of the Society of Industrial and Organizational Psychology, Atlanta, GA.

Clevenger, J., Pereira, G. M., Wiechmann, D., Schmitt, N., & Schmidt Harvey, V. (2001). Incremental validity of situational judgment tests. *Journal of Applied Psychology, 86*, 410–417.

Cronbach, L. J. (1949). *Essentials of psychological testing*. New York, NY: Harper & Row.

Cucina, J. M., Vasilopoulos, N. L., & Leaman, J. A. (2003, April). *The bandwidth-fidelity dilemma and situational judgment test validity*. Paper presented at the 18th Annual Conference of the Society of Industrial and Organizational Psychology, Orlando, FL.

Dalessio, A. T. (1994). Predicting insurance agent turnover using a video-based situational judgment test. *Journal of Business and Psychology, 9,* 23–32.

Devlin, S. E., Abrahams, N. M., & Edwards, J. E. (1992). Empirical keying of biographical data: Cross-validity as a function of scaling procedure and sample size. *Military Psychology, 4*, 119–136.

File, Q. W. (1943). *How Supervise? (Questionnaire Form B)*. New York, NY: The Psychological Corporation.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 41*, 237–358.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt & F. Ostendorf (Eds.), *Personality Psychology in Europe, Vol. 7* (pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84–96.

Greenacre, M. J. (2007). *Correspondence Analysis in Practice* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19–30.

Guilford, J. P., & Lacey, J. I. (1947). *Printed Classification Tests*. Army Air Forces Aviation Psychology Program, Report No. 5. Washington, DC: U.S. Government Printing Office.

Hanson, M. A., & Borman, W. C. (1995). *Development and construct validation of the situational judgment test.* ARI Research Note 95-34. Personnel Decisions Research Institutes, Inc.

Hanson, M. A., Borman, W. C., Mogilka, H. J., Manning, C., & Hedge, J. W. (1999). Computerized assessment of skill for a highly technical job. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 197–220). Mahwah, NJ: Erlbaum.

Hanson, M. A., Horgen, K. E., & Borman, W. C. (1998). *Situational judgment: An alternative approach to selection test development* [On-line]. Available: http://www.internationalmta.org/1998/9834d.html

Hogan, J. B. (1994). Empirical keying of background data measures. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 69–107). Palo Alto, CA: Consulting Psychologists Press.

Hough, L., & Paullin, C. (1994). Construct-oriented scale construction: The rational approach. In G. S. Stokes, M. D. Mumford, & W. A. Owens (Eds.), *Biodata handbook: Theory, research, and use of biographical information in selection and performance prediction* (pp. 109–145). Palo Alto: Consulting Psychologists Press.

Hunter, D. R. (2003). Measuring general aviation pilot judgment using a situational judgment technique. *The International Journal of Aviation Psychology, 13,* 373–386.

Jones, M. W., Dwight, S. A., & Nouryan, T. R. (1999, April). *Exploration of the construct validity of a situational judgment test used for managerial assessment*. Paper presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Kanning, U. P., Grewe, K., Hollenberg, S., & Hadouch, M. (2006). From the subjects' point of view – Reactions to different types of situational judgment items. *European Journal of Psychological Assessment, 22,* 168–176.

Kiers, H. A. L. (1991). Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika, 56,* 449–470.

Kihlstrom, J. F., & Cantor, N. (2000). Social intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (2nd ed., pp. 359–379). Cambridge, U.K.: Cambridge University Press.

Kleinmuntz, B. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin, 107*, 296–310.

Krokos, K. J., Meade, A. W., Cantwell, A. R., Pond, S. B., & Wilson, M. A. (2004, April). *Empirical keying of situational judgment tests: Rationale and some examples.* Paper presented at the 19th Annual Meeting of the Society for Industrial and Organizational Psychology.

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 65*, 422–427.

Latham, G. P., & Wexley, K. N. (1982). *Increasing productivity through performance appraisal.* Reading, MA: Addison-Wesley.

Legree, P. J., Psotka, J., Tremble, T., & Bourne, D. (2005). *Using consensus based measurement to assess emotional intelligence.* In R. Schulze & R. D. Roberts (Eds.), Emotional intelligence: An international handbook (pp. 155–179). Cambridge, MA: Hogrefe & Huber.

Lievens, F. (2000). Development of an empirical scoring scheme for situational inventories. *European Review of Applied Psychology, 50*, 117–124.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). The operational validity of a videobased situational judgment test for medical college admissions: Illustrating the importance of matching predictor and criterion construct domains. *Journal of Applied Psychology, 90*, 442–452.

Lievens, F., & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology, 86*, 1202–1222.

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational judgment tests: A review of recent research. *Personnel Review, 37*, 426–441.

Lievens, F., & Sackett, P. R. (2006). Video-based versus written situational judgment tests: A comparison in terms of predictive validity. *Journal of Applied Psychology, 91,* 1181–1188.

McDaniel, M. A., Hartman, N. S., & Grubb, W. L., III. (2003, April). *Situational judgment tests, knowledge, behavioral tendency, and validity: A meta-analysis.* Paper presented at the 18th Annual Conference of the Society for Industrial and Organizational Psychology. Orlando, FL.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L., III. (2007). Situational judgment tests, response instructions, and validity: A metaanalysis. *Personnel Psychology, 60,* 63–91.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Predicting job performance using situational judgment tests: A clarification of the literature. *Journal of Applied Psychology, 86,* 730–740.

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9,* 103–113.

McDaniel, M. A., Powell Yost, A., Ludwick, M. H., Hense, R. L., & Hartman, N. S. (2004, April). *Incremental validity of a situational judgment test.* Paper presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

McDaniel, M. A., & Whetzel, D. L. (2005). Situational judgment test research: Informing the debate on practical intelligence theory. *Intelligence, 33,* 515–525.

McDaniel, M. A., & Whetzel, D. L. (2007). Situational judgment tests. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement: Industrial psychology in human resources management* (pp. 235–257). Hillsdale, NJ: Erlbaum.

McDaniel, M. A., Whetzel, D. L., & Nguyen, N. T. (2006). *Situational judgment tests in personnel selection: A monograph for the International Personnel Management Association Assessment Council.* Alexandria, VA: International Personnel Management Assessment Council.

McElreath, J., & Vasilopoulos, N. L. (2002, April). *Situational judgment: Are most and least likely responses the same?* Paper presented at the 17th Annual Conference of the Society for Industrial and Organizational Psychology, Toronto, Canada.

McHenry, J. J., & Schmitt, N. (1994). Multimedia testing. In M. G. Rumsey, C. B. Walker & J. H. Harris (Eds.), *Personnel selection and classification* (pp. 192–232). Hillsdale, HJ: Erlbaum.

Meehl, P. E. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102*, 246–268.

Moss, F. A., Hunt, T., Omwake, K. T., & Ronning, M. M. (1927). *Social Intelligence Test.* Washington, D.C.: Center for Psychological Service.

Motowidlo, S. J., Borman, W., & Schmit, M. (1997). A theory of individual differences in task and contextual performance. *Human Performance, 10*, 71–83.

Motowidlo, S. J., Carter, G. W., Dunnette, M. D., Tippins, N., Werner, S., Burnett, J. R. et al. (1992). Studies of the structured behavioral interview. *Journal of Applied Psychology, 77,* 571–587.

Motowidlo, S. J., Diesch, A. C., & Jackson, H. L. (2003, April). *Using situational judgment format to measure personality characteristics.* Paper presented at the 18th annual conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75,* 640–647.

Motowidlo, S. J., Hanson, M. A., & Crafts, J. L. (1997). Low-fidelity simulations. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measure methods in industrial psychology* (pp. 241–260). Palo Alto, CA: Davis Black.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006a). Implicit policies about relations between personality traits and behavioral effectiveness in situational judgment items. *Journal of Applied Psychology, 91,* 749–761.

Motowidlo, S. J., Hooper, A. C., & Jackson, H. L. (2006b). A theoretical basis for situational judgment tests. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 57–81). Mahwah, NJ: Erlbaum.

Mumford, M. D. (1999). Construct validity and background data: Issues, abuses, and future directions. *Human Resource Management Review, 9*, 117–145.

Mumford, M. D., & Owens, W. A. (1987). Methodology review: Principles, procedures, and findings in the application of background data measures. *Applied Psychological Measurement, 11*, 1–31.

Mumford, M. D., & Whetzel, D. L. (1997). Background data. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology* (pp. 207–239). Palo Alto, CA: Davies-Black.

Mumford, T. V., Campion, M. A., & Morgeson, F. P. (2006). Situational judgment in work teams: A team role typology. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 319–343). Mahwah, NJ: Erlbaum.

Mumford, T. V., Van Iddekinge, C. H., Morgeson, F. P., & Campion, M. A. (2008). The team role test: Development and validation of a team role know-

ledge situational judgment test. *Journal of Applied Psychology, 93*, 250–267.

Nguyen, N. T., Biderman, M. D., & McDaniel, M. A. (2005). Effects of response instructions on faking a situational judgment test. *International Journal of Selection and Assessment, 13*, 250–260.

Nguyen, N. T., & McDaniel, M. A. (2003). Response instructions and racial differences in a situational judgment test. *Applied H.R.M. Research, 8*, 33–44.

O'Connell, M. S., Hartman, N. S., McDaniel, M. A., Grubb, W. L., & Lawrence, A. (2007). Incremental validity of situational judgment tests for task and contextual job performance. *International Journal of Selection and Assessment, 15*, 19–29.

Olson-Buchanan, J. B., Drasgow, F., Moberg, P. J., Mead, A. D., Keenan, P. A., & Donovan, M. A. (1998). Interactive video assessment of conflict resolution skills. *Personnel Psychology, 51,* 1–24.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89,* 187–207.

Paulhus, D. L. (1984). Two-component models of social desirable responding. *Journal of Personality and Social Psychology, 46*, 598–609.

Pereira, G. M., & Schmidt Harvey, V. (1999, April). *Situational judgment tests: Do they measure ability, personality or both*. Paper presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Phillips, J. F. (1992). Predicting sales skills. *Journal of Business and Psychology, 7*, 151–160.

Phillips, J. F. (1993). Predicting negotiation skills. *Journal of Business and Psychology, 7*, 403–411.

Ployhart, R. E. (2006). Staffing in the 21st century: New challenges and strategic opportunities. *Journal of Management, 32*, 868–897.

Ployhart, R. E., & Ehrhart, M. G. (2003). Be careful what you ask for: Effects of response instructions on the construct validity and reliability of situational judgment tests. *International Journal of Selection and Assessment, 11,* 1–16.

Ployhart, R. E., Porr, W. B., & Ryan, A. M. (n. d.). *Developing situational judgment tests in a service context: Exploring an alternative methodology.* Unpublished manuscript.

Ployhart, R. E., & Ryan, A. M. (2000, April). *Integrating personality tests with situational judgment tests for the prediction of customer service performance*. Paper presented at the 16th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Porr, W. B., & Ployhart, R. E. (2004, April). *The validity of empirically and construct-oriented situational judgment tests*. Paper presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Reynolds, D. H., Winter, J. L., & Scott, D. R. (1999, April). *Development, validation, and translation of a professional-level situational judgment inventory.* Paper presented at the 14th Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology, 85*, 880–887.

Rosen, N. A. (1961). How Supervise?–1943–1960. *Personnel Psychology, 14*, 87–99.

Sacco, J. M., Schmidt, D. B., & Rogg, K. L. (2000, April). *Using readability statistics and reading comprehension scores to predict situational judgment test performance, black-white differences, and validity.* Paper presented at the 15th Annual Conference of the Society for Industrial and Organizational Psychology, New Orleans, LA.

Salgado, J. F., Viswesvaran, C., & Ones, D. S. (2001). Predictors used for personnel selection: An overview of constructs, methods, and techniques. In N. R. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work & organizational psychology: Vol. 1* (pp. 165–199). London: Sage.

Schmidt, F. L., & Hunter, J. E. (1993). Tacit knowledge, practical intelligence, general mental ability, and job knowledge. *Current Directions in Psychological Science, 2*, 8–9.

Schmitt, N., & Chan, D. (2006). Situational judgment tests: Method or construct? In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 135–156). Mahwah, NJ:

Erlbaum.

Shoda, Y., Mischel, W., & Wright, J. C. (1993). The role of situational demands and cognitive competencies in behavior organization and personality coherence. *Journal of Personality and Social Psychology, 65*, 1023–1035.

Smith, K. C., & McDaniel, M. A. (1998, April). *Criterion and construct validity evidence for a situational judgment measure*. Paper presented at the 13th Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

Stemler, S. E., Elliot, J., Grigorenko, E. G., & Sternberg, R. J. (2006). There's more to teaching than instruction: Seven strategies for dealing with the practical side of teaching. *Educational Studies, 32*, 85–102.

Stemler, S. E., & Sternberg, R. J. (2006). Using situational judgment tests to measure practical intelligence. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 107–131). Mahwah, NJ: Erlbaum.

Sternberg, R. J. (1998). *Erfolgsintelligenz. Warum wir mehr brauchen als EQ + IQ.* München: Lichtenberg.

Sternberg, R. J. (1999). The theory of successful intelligence. *Review of General Psychology, 3*, 292–316.

Sternberg, R. J., Wagner, R. K., Williams, W. M., & Horwath, J. A. (1995). Testing common sense. *American Psychologist, 50*, 912–926.

Stevens, M. J., & Campion, M. A. (1999). Staffing work teams: Development and validation of a selection test for teamwork. *Journal of Management, 25,* 207–228.

Thornton, G. C., III, & Byham, W. C. (1982). *Assessment centers and managerial performance*. New York, NY: Academic Press.

Trippe, M. D. (2002). *An evaluation of construct validity of situational judgment tests.* Unpublished master's thesis, Polytechnic Institute and State University of Virginia at Blacksburg.

Trippe, M. D., & Foti, R. J. (2003, April). *An evaluation of the construct validity of situational judgment tests.* Paper presented at the 18th annual conference of the Society for Industrial and Organizational Psychology, Orlando, FL.

Van Vianen, A. E. M., Taris, R., Scholten, E., & Schinkel, S. (2004). Perceived fairness in personnel selection: Determinants and outcomes in different stages of the assessment procedure. *International Journal of Selection and Assessment, 12*, 149–159.

Wagner, R. K. (1987). Tacit knowledge in everyday intelligent behavior. *Journal of Personality and Social Psychology, 52*, 1236–1247.

Wagner, R. K., & Sternberg, R. J. (1985). Practical intelligence in real world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49*, 436–458.

Weekley, J. A., Harding, R., Creglow, A., & Ployhart, R. E. (2004, April). *Scoring situational judgment tests: Does the middle matter?* Paper presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL.

Weekley, J. A., & Jones, C. (1997). Video-based situational testing. *Personnel Psychology, 50,* 25–49.

Weekley, J. A., & Jones, C. (1999). Further studies of situational tests. *Personnel Psychology, 52,* 679–700.

Weekley, J. A., & Ployhart, R. E. (Eds.). (2006). *Situational judgment tests: Theory, measurement, and application*. Mahwah, NJ: Erlbaum.

Weekley, J. A., Ployhart, R. E., & Harold, C. M. (2004). Personality and situational judgment tests across applicant and incumbent settings: An examination of validity, measurement, and subgroup differences. *Human Performance, 17*, 433–461.

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and scoring. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, management, and application* (pp. 157 – 182). Mahwah, NJ: Erlbaum.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*, 372–376.

Whetzel, D. L., McDaniel, M. A., & Nguyen, N. T. (2008). Subgroup differences in situational judgment test performance: A meta-analysis. *Human Performance, 21*, 291–309.

Wright, J. C., & Mischel, W. (1987). A conditional approach to dispositional constructs: The local predictability of social behavior. *Journal of Personality and Social Psychology, 53*, 1159–1177.

3. Der Act Frequency Approach

3.1 Grundannahmen beim Act Frequency Approach

Die Grundlage für die Entwicklung von SJTs liefern *Subject Matter Experts* – zum Beispiel langjährige Mitarbeiter oder Personalfachleute – welche für die zu besetzende Arbeitsstelle erfolgskritische Situationen beschreiben und anschliessend die dafür repräsentativsten auswählen. Gilt es, dieses Vorgehen für die Entwicklung eines Instruments zur Erfassung von Persönlichkeitseigenschaften einzusetzen, so bildet das Ausgangsmaterial die Beschreibung prototypischer Handlungen, welche die zu messende Persönlichkeitseigenschaft bestmöglich charakterisieren. Für die Persönlichkeitsdimension „Gewissenhaftigkeit" wäre dies zum Beispiel die Aussage „Er überprüfte die Daten und seine Berechnungen mehrmals, bevor er die Statistiken den Zeitungen zur Veröffentlichung zustellte."

Ein für die systematische Generierung solcher für verschiedene Persönlichkeitseigenschaften prototypischen Handlungen geeignetes Vorgehen entwickelten Buss und Craik (1980, 1981, 1983a, 1984, 1989): den *Act Frequency Approach* (AFA; auch Handlungs-Häufigkeits-Ansatz genannt). Dieser basiert auf der Überlegung, dass sich Handlungen zu Kategorien gruppieren lassen und so als Manifestationen derselben Persönlichkeitseigenschaft angesehen werden können. Für Buss und Craik stellen somit Persönlichkeitseigenschaften nichts weiter dar als mehr oder weniger klar abgrenzbare Kategorien von Handlungen („*dispositions as natural cognitive categories of acts*"; z. B. Buss & Craik, 1983b, S. 396). Als dominante Person kann gemäss dieses Verständnisses der Persönlichkeit bezeichnet werden, wer in verschiedenen Situationen mehr dominantes Verhalten gegenüber Mitmenschen zeigt als andere Personen dies tun.

Buss und Craik (z. B. 1983a) setzten in ihren Studien zum AFA folgendes Vorgehen ein: Die Studienleiter fordern die Probanden auf, an drei Personen in ihrem Bekanntenkreis zu denken, bei denen die gefragte Eigenschaft besonders ausgeprägt ist. Sie sollen daraufhin in der Vergangenheit gezeigtes Verhalten dieser Personen beschreiben, welches diese Eigenschaft sehr typisch charakterisiert. Die auf diese Weise generierten Listen mit Verhaltensweisen (Acts) legen die Studienleiter einer zweiten Gruppe von Probanden vor, mit dem Auftrag, die gesammelten Acts nach deren Prototypizität hinsichtlich der gefragten Eigenschaft einzustufen. So schätzten zum Beispiel die Probanden in der von Buss und Craik (1980) durchgeführten Studie aus einer Liste von 100 Acts zur Persönlichkeitseigenschaft Dominanz „Er/Sie erteilte Anweisungen, um die Gruppe zu orga-

nisieren." oder „Er/Sie beeinflusste das Ergebnis eines Meetings, ohne dass die anderen dies bemerkten." als besonders prototypisch ein. Für den letzten Schritt der Entwicklung eines Messinstruments werden die prototypischsten Acts ausgewählt und mit einer dreistufigen Antwortskala versehen, auf welcher die befragten Personen angeben können, ob sie die Handlung in ihrem Leben selten, manchmal oder oft ausgeführt haben. Gemäss einer Grundannahme des AFA gilt nun, dass die Anzahl der Acts, die eine Person in der Vergangenheit gezeigt hat, ein Mass für die Ausprägung der zu messenden Persönlichkeitseigenschaft darstellt.

Moser (1989) charakterisiert die dem AFA zu Grunde liegenden Annahmen mittels vier Statements:

1. Persönlichkeitseigenschaften stellen nichts weiter dar als in Klassen zusammengefasste Verhaltensweisen: „The act frequency approach views dispositions as cognitive categories of acts that serve to summarize general trends in behavior ... [and] as appropriate predictions of future trends in conduct" (Buss & Craik, 1985, S. 936).
2. Die verschiedene Persönlichkeitseigenschaften repräsentierenden Verhaltensklassen lassen sich nicht scharf voneinander abgrenzen – nach Zadeh, Fu, Tanaka und Shimura (1975) lassen sie sich als *fuzzy sets* bezeichnen – sondern gehen fliessend ineinander über, da die Prototypizität der sie konstituierenden Verhaltensweisen vom Zentrum einer Kategorie zu deren Peripherie hin stetig abnimmt.
3. Mit dem AFA lassen sich gemäss Buss und Craig (1985) mehrere Kriterien formulieren, die der Einstufung der theoretischen und empirischen Bedeutung von Persönlichkeitseigenschaften dienen: (1) Eindeutigkeit, Bedeutsamkeit und Angemessenheit des Umfangs der Verhaltensklasse, (2) Unterscheidungskraft gegenüber anderen Verhaltensklassen, (3) übereinstimmende Beurteilungen über die Zugehörigkeit von Acts zu einer Verhaltensklasse und deren Prototypizität, (4) Zeitstabilität der Persönlichkeitseigenschaft, (5) beobachtbare interindividuelle Unterschiede in den Acts einer Persönlichkeitseigenschaft und (6) Auftretenshäufigkeit der Acts einer Persönlichkeitseigenschaft.
4. Mit dem AFA lassen sich individuelle Persönlichkeitsunterscheide untersuchen, indem die korrespondierenden Acts ausgezählt werden: „The act frequency approach to dispositional assessment does provide, in principle, an absolute metric, because it entails a true zero point: when no acts within the dispositional category are manifested for the period of observations" (Buss & Craik, 1984, S. 246).

Nach dieser Kurzdarstellung des AFA gehe ich in den nachfolgenden Kapiteln vertieft darauf ein, indem ich beschreibe, auf welche theoretischen Konzepte sich Buss und Craik abstützen, welche Phasen sie bei der Durchführung des AFA unterscheiden und welche Kritik zum AFA geäussert wurde. Den Abschluss bildet die vollständige und umfassende Darstellung der Durchführung des AFA anhand eines konkreten Beispiels.

3.2 Die Person-Situations-Debatte als Ausgangspunkt des Act Frequency Approachs

Buss und Craik (1980) formulierten den AFA als einen Beitrag zur Person-Situations-Debatte – auch als Interaktionismus-Debatte (Magnusson & Endler, 1977) oder Konsistenzdebatte (Mischel & Peake, 1982) bezeichnet –, welche Mischel 1968 durch seine Publikation „*Personality and Assessment*" auslöste. Darin kritisierte er die einseitig am Trait-Konzept verhaftete Persönlichkeitsforschung. Dieses Konzept postuliert, dass die – verborgenen – Persönlichkeitseigenschaften eine stabile Beziehung zwischen Situationen und Reaktionen des Individuums erzeugen. Kritik am Eigenschaftsparadigma war zu diesem Zeitpunkt nicht neu: Schon 1935 formulierte Lorge einen Einwand dazu und in den 60er Jahren begannen sich kritische Voten zu häufen (z. B. Bandura & Walters, 1963; Wallace, 1966) – Mischel trat mit seinem Buch jedoch eine Lawine los. Er führte darin verschiedene Studien über den Zusammenhang zwischen Persönlichkeit und Verhalten auf, in welchen die Autoren selten von Korrelationen über .30 berichteten – einem Wert, welcher Mischel (1968) als *„personality coefficient*" bezeichnete. Es schien so zu sein, dass sich die Annahme der transsituativen Stabilität von Persönlichkeitseigenschaften nicht bestätigen liess und dass die Wahrscheinlichkeit einer konsistenten Reaktion einer Person umso geringer wird, je unterschiedlicher die auslösenden Situationen sind. Die Verfechter der Trait-Theorie führten dies jedoch auf die Ungenauigkeit der Messtechnik und die daraus resultierenden Messfehler zurück.

Gordon W. Allport (1937) – welcher zusammen mit William Stern, Hans J. Eysenck und Raymond B. Cattell zu den Begründern der Trait-Theorie der Persönlichkeit gehört (Asendorpf, 1999; Pervin & John, 2001) – definiert Persönlichkeits-Traits als breit angelegte, basale Dispositionen, welche das Verhalten des Individuums über verschiedene Situationen hinweg beeinflussen und so zu Konsistenz im Verhalten führen. Mischel (1973) äusserte sich dazu wie folgt:

> These dispositions are not directly observed but are inferred from behavioral signs (trait indicators), either directly or indirectly … Guided by this assumption, personality research has been a quest for such underlying broad dimensions, for basic factors, or for pervasive motives, or for characteristic life styles. In personality assessment the trait assumptions regarding structure are seen in the existence of hundreds of tests designed to infer dispositions and almost none to measure situations. (S. 253)

Mischel forderte daraufhin, dass die Persönlichkeitsdiagnostik um Verfahren zu ergänzen ist, welche das Verhalten in Situationen erfassen. Wie wir sehen werden, stellt die Sammlung von konkreten Verhaltensweisen in realen Situationen das Kernstück des AFA dar: „Neither personality scales nor trait ratings constitute the basic measure of the frequency concept of disposition; rather the frequency concept involves sampling behavior by monitoring the relative frequency of specific acts within an appropriate response category over an array of occasions" (Buss & Craik, 1980, S. 380).

3.3 Theoretische Grundlagen des Act Frequency Approachs: Auftretenshäufigkeit und Prototypizität von Verhaltensweisen

Buss und Craik (1980) verbanden in ihrem Act Frequency Approach das *„frequency concept of dispositions"* von Alston (1975) mit dem Prototypenansatz von Rosch (1975, 1978)[1]. Diese beiden Konzepte stellen auch die zentralen Vorgehensschritte bei der Anwendung des AFA dar: Zuerst bittet man Probanden, konkrete, im Zusammenhang mit einer bestimmten Persönlichkeitseigenschaft stehende, Handlungen zu formulieren und legt diese dann einer zweiten Probandengruppe vor, um deren Prototypizität hinsichtlich der in Frage stehenden Persönlichkeitseigenschaft einzustufen.

Alston's *„frequency concept of dispositions"* – vom Autor selbst ursprünglich als *„S-R frequency disposition model"* bezeichnet (Alston, 1975, S. 21) – besagt, dass wir einer Person eine bestimmte Persönlichkeitseigenschaft zuschreiben, wenn diese in einer repräsentativen Auswahl von Situationen eine grosse Anzahl entsprechender Verhaltensweisen zeigt. Die Idee der Festlegung einer Persönlichkeitseigenschaft über die summarische Betrachtung von in ver-

[1] Buss und Craik (1980, 1981, 1983a) nennen als Grundlage des AFA zusätzlich das Konzept der „hypothetischen Annahmen" von Ryle (1949). Da sie sich in den späteren Schriften (1983b, 1983c, 1986a, 1986b) nicht mehr darauf beziehen, gehe ich an dieser Stelle nicht näher darauf ein.

schiedenen Situationen gezeigten Verhaltens beschrieb schon 20 Jahre früher Hampshire (1953):

> A statement which refers to a disposition [as "X is honest"] … summarises what tends to happen. … One can properly claim to know that someone has a certain disposition when (a) one has had occasion for prolonged and continuous observation of the conduct and calculations of the person in question, and (b) when one can quote many incidents in which the disposition manifested itself ... (S. 5-6)

Jaccard (1974) nutzte diese Überlegungen, um eine Persönlichkeitsskala zu entwickeln, welche sich aus Situationsschilderungen zusammensetzt. Dies stellt nicht per se eine Pionierleistung dar – mit dem *S-R Inventory of Anxiousness* haben zum Beispiel Endler, Hunt und Rosenstein schon 1962 ein situatives Persönlichkeitsinventar entwickelt –, Jaccard unternahm seine Studien aber vor dem Hintergrund der Person-Situations-Debatte und seine Vorgehensweise floss als zentraler Vorgehensschritt in den AFA von Buss und Craik ein. Jaccard stützte sich dabei auf Fishbein (1972) ab, welcher drei verschiedene Kriterien zur Verhaltenseinstufung beschreibt: das *„single act, single observation criterion"* (zeigt das Individuum ein bestimmtes Verhalten in einer bestimmten Situation?), das *„single act, repeated observation criterion"* (zeigt das Individuum ein bestimmtes Verhalten zu verschiedenen Zeitpunkten?) und das *„multiple act criterion*" (zeigt das Individuum ähnliches Verhalten in unterschiedlichen Situationen?).

Jaccard legte seiner Studie dementsprechend folgende Überlegung zu Grunde: Wenn eine Persönlichkeitseigenschaft die allgemeine Ausprägung des dieser Eigenschaft zugeordneten Verhaltens steuert, jedoch nicht ein spezifisches Verhalten per se, so müsste ein Mass dieser Persönlichkeitseigenschaft höher mit dem *multiple act criterion* korrelieren als mit dem *single act criterion*. Er wählte die Persönlichkeitseigenschaft Dominanz und liess dazu von 22 Studentinnen das *multiple act criterion* generieren, indem diese die Aufgabe bekamen, „to think of a female college student or friend who you consider to be dominant and to write down five specific behaviors that you think this person has performed that would be related to her being dominant" (Jaccard, 1974, S. 361). Aus den so generierten Acts wählte er 40 aus, die er 45 Studienteilnehmern vorlegte, welche jeweils angeben mussten, ob sie dieses Verhalten in der Vergangenheit schon gezeigt hatten. Diese Angaben korrelierte er mit den Ergebnissen aus Dominanzskalen von zwei Persönlichkeitsinventaren (dem California Psychological Inventory und der Personality Research Form) und konnte so seine Hypothese bestätigen: Zwischen den Skalen und dem *multiple act criterion* zeigten sich Korrelationen zwischen r = .51 und .64, die mittleren Korrelationen zwischen den einzelnen

Acts und den Skalen betrug jedoch lediglich r = .17 bis .20 (siehe auch Buss & Craik, 1981).

Jaccard schlug vor, dass bei der Entwicklung von *multiple act criteria* ähnlich vorzugehen ist wie bei herkömmlichen Persönlichkeitsinventaren, indem man die Verhaltensbeispiele als Items einer Skala ansieht und in einem Score zusammenfasst. Weiter soll der Fragebogenentwickler bei der Itemselektion nicht intuitiv oder zufällig vorgehen, sondern die beiden Konzepte der klassischen Testtheorie „Trennschärfe" und „Schwierigkeit" anwenden: Der Act muss eindeutig einer Persönlichkeitseigenschaft zuordenbar sein (Kriterium *„behavior ambiguity*") und Acts, welche beinahe alle oder keiner der befragten Personen schon einmal gezeigt haben, sind auszuschliessen (Kriterium *„base rate*"; Jaccard, 1974, S. 363). Krüger und Amelang (1995) wandten das zweite Itemselektionskriterium, die Basisrate, bei ihrer Skalenkonstruktion an, ebenso Buss und Craik (1985), jedoch unter der Bezeichnung *„endorsement*". Beide Male bezeichneten damit die Autoren den Prozentsatz an Personen, die angegeben haben, dass sie diesen Act mindestens einmal ausgeführt haben.

Indem Jaccard auf die Bedeutung der Auswahl „angemessener" und eindeutiger Situationen für die Bildung valider situationsbasierter Persönlichkeitsskalen hinweist, legt er auch den Grundstein für die Überlegungen zum zweiten zentralen Vorgehensschritt des AFA von Buss und Craik, der Einstufung der Prototypizität der Acts (Rosch & Mervis, 1975). Damit lassen sich die Acts bezüglich ihrer Zugehörigkeit zu einer bestimmten Kategorie in eine Rangreihenfolge bringen, welche von zentralen, prototypischen Beispielen dieser Kategorie zu weniger prototypischen, peripheren reicht. Rosch und Mervis (1975) haben in ihrer Schrift das Prototypen-Konzept am Beispiel von Hunderassen erläutert: Wir alle haben eine Vorstellung davon, wie ein „richtiger", ein „hundiger" Hund ausschaut. Und so ist für die meisten von uns ein Schäferhund ein eher typischer Hund und ein Pekinese ein eher weniger typischer. Cantor und Mischel (1977) gehören zu den ersten, welche die Idee des Prototypenratings auf die Erforschung der Persönlichkeit anwendeten, indem sie 200 Adjektive zu Persönlichkeitseigenschaften hinsichtlich deren Zusammenhangs zu den Konzepten introvertiert und extravertiert einstufen liessen.

3.4 Die Phasen des Act Frequency Approachs

Buss und Craik (1984, 1986a, 1986b) unterteilen die Durchführung des AFA in vier Phasen: Erstens das Sammeln der Acts, zweitens das Einschätzen der Prototypizität, drittens die Mehrfachsortierung und viertens das Erheben der Auftretenshäufigkeit der einzelnen Acts. Nachfolgend stelle ich die einzelnen Phasen anhand von Zitaten der Anweisungen von Buss und Craik dar:

1. Sammeln der Acts (act nominations)

Zu Beginn des AFA steht die Generierung einer umfassenden Liste mit konkreten Situationsschilderungen, indem man Probanden den Auftrag erteilt, zur interessierenden Verhaltensdimension passendes Verhalten (Acts) zu schildern. Die erste diesbezügliche Datenerhebung von Buss und Craik (1980) stellte eine Replikation der Studie von Jaccard (1974) dar, wobei sie auch dessen Anweisung übernahmen:

> Think of three of the most dominant people you know of your own gender. With these dominant individuals in mind, write down five acts or behaviors they might perform that would reflect or exemplify their dominance. Now think of three people you know of the opposite sex and list five acts or behaviors that would reflect or exemplify their dominance. (Buss & Craik, 1980, S. 381)

Angleitner und Demtröder (1988) kritisierten diese Anweisung, da sie in ihrer Datenerhebung dazu führte, dass die Probanden wenig konkretes Verhalten schilderten, sondern vielmehr Eindrücke oder Namen für einzelne Persönlichkeitszüge. Später modifizierten Buss und Craik (1989, siehe auch Buss, 1988) auf Grund ähnlicher Erfahrungen diese Anweisung, um zu verhindern, dass die Probanden zu generelle Situationsschilderungen liefern oder Verhaltenszuschreibungen (z. B. „Sie verhielt sich dominant.“) nennen:

> In this study, we are interested in the things people do to get ahead: How do people climb, elevate, jockey, or defend positions in the status or dominance hierarchy? Please be specific: We are interested in specific acts or behaviors. One should be able to answer the following questions about each of your act nominations: Have you ever performed this act? If so, how often have you performed it? Please think of specific people you know (including yourself) of your own sex, and write down five acts that they have performed to get ahead in status or dominance hierarchies. Now think of people you know of the opposite sex and write down five acts or

> behaviors that they have performed to get ahead in status or dominance hierarchies. (S. 24)

Buss und Craik sichten jeweils anschliessend an diesen Arbeitsschritt die Listen mit den generierten Acts, sortieren völlig unpassende, nicht regelkonform formulierte (zu vage, kein Verhalten, generelle Tendenz, Häufigkeitsangaben) und redundante aus und überarbeiten die restlichen hinsichtlich Grammatik und Satzbau. Die so erarbeiteten Listen mit Verhaltensweisen stellen das Ausgangsmaterial für die nachfolgenden Phasen zur Identifikation der internen Struktur der Kategorien und zur Auftretenshäufigkeit der Acts dar.

2. Einschätzen der Prototypizität (prototypicality ratings resp. single prototypicality rating)

Bei der zweiten Phase, der Einschätzung der Prototypizität der generierten Acts, lehnten sich Buss und Craik stark an die Studie von Rosch und Mervis (1975) an und setzten dieselbe Instruktion und dieselbe Antwortskala ein:

> This study has to do with what we have in mind when we use words which refer to categories. Let's take the word *red* as an example. Close your eyes and imagine a true red. Now imagine an orangish red ... imagine a purple red. Although you might still name the orange-red and the purple-red with the term *red*, they are not as good examples of red (as clear cases of what *red* refers to) as the clear „true" red. In short, some reds are „redder" than others.
>
> In this study you are asked to judge how good an example of a category various instances of the category are. The category is dominance. Below are listed 100 acts. You are to rate how good an example of that category each act is on a 7-point scale. A „7" means that you feel the act is a very good example of your idea of what dominance is; a „1" means you feel the act fits very poorly with your idea of what dominance is (or is not a member of that category at all). A „4" means you feel the act fits moderately well. Use the other numbers of the 7-point scale to indicate intermediate judgments. (Buss & Craik, 1980, S. 382–383)

Es zeigte sich, dass hoch spezifische und gut beobachtbare Acts eine hohe Prototypizität zugewiesen bekamen (Block, 1989). Verschiedene Forschergruppen (z. B. Amelang, Herboth & Oefner, 1991; Buss & Craik, 1980; Krüger & Amelang, 1995) konnten zudem empirisch nachweisen, dass hochprototypische Acts höher mit externen Kriterien korrelieren als niedrigprototypische, wobei die gefundenen Unterschiede zum Teil beträchtlich sind: In der von Buss und Craik

(1980) durchgeführten Studie korrelieren die 25% hochprototypischsten Acts zur Dimension Dominanz mit r = .48, die 25% niedrigprototypischen r = .14 mit der entsprechenden Skala aus dem CPI (Gough, 1957). Diese Ergebnisse liefern eine Bestätigung des Funktionierens des im AFA integrierten Prototypenansatzes, welcher seinerseits als ein empirisch bestätigtes Vorgehen zur Beschreibung der internen Struktur von Verhaltenskategorien angesehen wird: „Die sehr hohe Übereinstimmung zwischen den Versuchspersonen in der Abgabe der Prototypen-Urteile kann als Beleg für die allgemeine Verbreitung von Vorstellungen über Eigenschaften und die sie konstituierenden Verhaltensweisen gewertet werden" (Amelang & Bartussek, 1990, S. 67). Wie ich weiter unten darstelle, ist jedoch die Höhe der Beurteilerübereinstimmung und somit der Prototypenansatz als nützliches Mittel in der Persönlichkeitsforschung umstritten.

3. Mehrfachsortierung (multiple dispositional act sorting)

Schon kurze Zeit nach der Veröffentlichung des AFA, erweiterten Buss und Craik (1984) das Prototypenrating, indem sie die Mehrfachsortierung (*multiple dispositional act sorting*) einführten. Dies auf Grund der Erfahrung, dass die Probanden beim Sammeln der Acts zum Teil Verhaltensweisen nannten, die besser zu einer anderen Kategorie passen, als zur aktuell bearbeiteten. Zudem kann ein Act auch prototypisch für zwei oder mehrere Kategorien sein, vor allem wenn die Kategorien semantisch nahe beieinander liegen. Buss und Craik (1986a) schreiben dazu:

> Although prototypicality ratings yield simple and direct indices of the differential status of acts, they undoubtedly underestimate the complexity of the multiple constructs that may be used to interpret each act. In particular, some acts may be subsumed by more than one dispositional construct, especially if they fall toward the periphery of categories. (S. 147)

Das oben dargestellte Prototypenrating lässt es jedoch nicht zu, unpassende Acts einer anderen Kategorie zuzuordnen. So führten Buss und Craik diese zusätzliche Phase ein, in welcher sie die Probanden in einem ersten Arbeitsschritt auffordern, die Acts vorgegebenen Kategorien zuzuordnen, wobei diese ein Act auch zwei oder mehreren Kategorien zuordnen dürfen. Zudem ist es den Probanden auch erlaubt, zusätzliche Kategorien zu bilden, wenn sie dies für notwendig erachteten. Im zweiten Arbeitsschritt stufen die Probanden die Prototypizität jedes Acts für jede der zugeteilten Kategorien ein.

Dieses Vorgehen kritisieren Angleitner und Demtröder (1988), weil es dazu führt, dass nur diejenigen Probanden die Prototypizität eines Acts in einer oder

mehreren zusätzlichen Kategorie einstufen können, welche den Act vorgängig auch diesen Kategorien zugeordnet haben. Die Einstufung der Prototypizität der verschiedenen Acts basiert dann auf einer unterschiedlichen Anzahl von Ratings. Sie schlagen deshalb vor, dass alle Rater für jeden Act die Prototypizität für jede Kategorie einstufen (siehe Angleitner, Buss & Demtröder, 1990).

Die Aussage von Buss und Craik (1989, S. 24), dass „the notion of simple single category membership is not applicable" und die daraus abgeleitete Einführung der Mehrfachsortierung verstösst gegen die Forderung von Jaccard (1974), dass der Act zwecks Verhinderung der *„behavior ambiguity*" eindeutig einer Persönlichkeitseigenschaft zuordenbar sein muss. Buss und Craik (1989) sehen in ihrem Vorgehen jedoch keinen Widerspruch zum Prototypenansatz, weil sich zum Beispiel auch Alltagsgegenstände verschiedenen Kategorien zuordnen lassen (z. B. Rosch, 1975). Buss und Craik (1983) nehmen an, dass die Korrelation zwischen zwei Verhaltenskategorien auf Acts zurückzuführen ist, welche in beiden Verhaltenskategorien auftreten und damit zur inhaltlichen Überlappung dieser beiden Kategorien beitragen.

4. Erheben der Auftretenshäufigkeit der Acts (assessing act performance)

Bei der Vorlage einer AFA-Skala, müssen die Probanden angeben, ob sie das geschilderte Verhalten in einer vorgegebenen Zeitperiode in der Vergangenheit schon gezeigt haben oder nicht und wenn ja, wie häufig (selten, ab und zu, häufig; Buss & Craik, 1980, 1983c). Smid, Douma, Van Lenthe und Ranchor (1988) wandelten diese Anweisung ab, indem sie die Probanden aufforderten anzugeben, wie hoch sie die Wahrscheinlichkeit einschätzen, das entsprechende Verhalten in der geschilderten Situation zu zeigen. Damit umgehen sie das Problem der grossen Variation in der Auftretenshäufigkeit verschiedener Acts (z. B. Buss & Craik, 1985; siehe auch nächstes Kapitel). Auch Höft (2002) und Muck, Höft, Hell und Schuler (2006) wandten diese Anweisung an und boten den Probanden eine fünfstufige Likert-Skala mit der Skalenverankerungen „sehr unwahrscheinlich" bis „sehr wahrscheinlich". Bei Johnson und Lecci (2003) mussten die Studienteilnehmer auf einer vierstufigen Skala, welche von „strongly disagree" bis „strongly agree" reicht, angeben, wie sie generell denken, fühlen oder handeln. Buss und Craik (1989, S. 26) kritisierten diese Abwandlungen jedoch scharf: „We do not know much about the merits of a procedure that asks respondents to make self-reflections about hypothetical conditions, but these estimates cannot be considered act frequency assessments."

Neben diesem von Buss und Craik (1984) als *„retrospective self-recoding of act frequencies*" genannten Vorgehen, schlagen sie zusätzlich noch die Durch-

führung einer Fremdbeurteilung vor, das *„retrospective observer recording of act frequencies*". Ausgewertet wird in beiden Fällen jeweils die Basisrate des Acts, das heisst wie häufig – in Prozent – die Probanden genannt haben, dass sie dieses Verhalten gezeigt haben. Die einzelnen Acts pro Verhaltenskategorie lassen sich zudem zu einer Skala zusammenfassen (*multiple-act index* resp. *acttrend index*). Angleitner, Buss und Demtröder (1990) konnten aufzeigen, dass die Auftretenshäufigkeit eines Acts unabhängig von dessen Prototypizität ist.

Buss und Craik (1983c; siehe auch Block, 1989) führten im Zusammenhang mit dem AFA noch die Begriffe *act density*, *act extensity* und *act bipolarity* ein. *Act density* bezieht sich auf die Anzahl signifikanter Korrelationen, welche eine überprüfte Persönlichkeitsskala mit den jeweiligen Acts der korrespondierenden Verhaltenskategorie hat. Damit lässt sich gemäss den Autoren die Bandbreite der Persönlichkeitsskala innerhalb der Verhaltenskategorie bestimmen, was sie mit den Konzepten der *content saturation* (Jackson, 1971) und der *bandwidth* (Cronbach & Gleser, 1965) vergleichen. Bezug nehmend auf das Circumplex-Modell der Persönlichkeit von Wiggins (1979), beschreibt *act bipolarity* die Anzahl der bedeutsamen Korrelationen von Acts einer Verhaltenskategorie mit der auf dem Circumplex-Modell entgegengesetzten Kategorie. *Act extensity* beschreibt die Korrelationen mit allen anderen Kategorien ausser der auf dem Circumplex-Modell entgegengesetzten. Buss und Craik (1985) führen zudem sechs Kriterien auf, welche Hinweise auf die empirische und theoretische Bedeutung einer Verhaltenskategorie liefern:

1. Anzahl, Spannbreite und Bedeutsamkeit der Acts der Verhaltenskategorie (*category volume*; Buss & Craik, 1983a).
2. Grad der Überlappung zwischen den Verhaltenskategorien, bedingt durch mehreren Kategorien zugeteilte Acts.
3. Grad der Beurteiler-Übereinstimmung bezüglich der Zugehörigkeit und der Prototypizität der Acts.
4. Ausprägung der zeitlichen Stabilität der Auftretenshäufigkeit der hochprototypischen Acts einer Kategorie.
5. Grösse der interindividuellen Differenz der Auftretenshäufigkeit der Acts.
6. Höhe der Basisrate der zu einem Wert pro Kategorie zusammengefassten Acts.

3.5 Kritik am Act Frequency Approach

Einige Wissenschaftler auf dem Gebiet der Persönlichkeitsforschung erkannten das Potenzial des AFA und setzten dies erkenntnisgewinnbringend um. So wies zum Beispiel Borkenau (1986) nach, dass die Interkorrelationen zwischen verschiedenen Persönlichkeitsdimensionen auf Strukturen in der Sprache *und* im Verhalten zurückzuführen sind. Einen wichtigen Beitrag liefert der AFA bei der Erforschung und somit dem besseren Verständnis von Persönlichkeitseigenschaften: Peterson (1993) zum Beispiel schloss aus den genannten Acts, dass Hilflosigkeitsverhalten häufig in Zusammenhang mit sozialen Interaktionen auftritt, wie etwa „Lass andere die Entscheidung treffen." oder „Lass dich von anderen Menschen unterstützen.". Botwin und Buss (1989) konnten das Big-Five-Modell anhand einer Sammlung von Acts replizieren, jedoch erst wenn sie vor der Durchführung der Faktorenanalyse bei den Acts das generelle Aktivitätsniveau, das heisst die Anzahl berichteter Acts pro Proband, herauspartialisierten. Andere Studien befassen sich mit dem interkulturellen Vergleich der konzeptionellen und der manifesten Struktur der Act-Dimensionen (z. B. Angleitner, Buss & Demtröder, 1990; Church, Katigbak, Miramontes, Del Prado & Cabrera, 2007; Willmann, Feldt & Amelang, 1997). Neben der Erforschung der Persönlichkeit, setzten einige Forscher den AFA auch zur Analyse und zur besseren Beschreibung psychiatrischer Krankheitsbilder (Buss & Craik, 1986; Shopshire & Craik, 1996; Sprock, 2000) oder des Verhaltens in Organisationen (z. B. Allen, 1993; Cinite, Duxbury & Higgins, 2009; Cooper, Dyke & Kay, 1990; Szamosi & Duxbury, 2002) ein. Allgemein scheint der AFA besonders dort effektiv zu sein und zu guten Ergebnissen zu führen, wo sich das interessierende Konstrukt theoretisch schlecht erfassen lässt, da sich anhand der Acts sehr genau beschreiben lässt, was die Leute unter diesem Konzept verstehen (z. B. Cinite et al., 2009; Cooper et al., 1990; siehe auch nachfolgendes Kapitel).

Insgesamt stiess der AFA bei den Persönlichkeitsforschern jedoch auf ein eher geringes Interesse. Auch in Lehrbüchern zur Persönlichkeitspsychologie und zu psychologischer Diagnostik taucht der AFA nur bei Autoren auf, welche diesen selbst schon eingesetzt haben (z. B. Amelang, Bartussek, Stemmler & Hagemann, 2006; Amelang & Schmidt-Atzert, 2006; Larsen & Buss, 2001). Die Literaturrecherche[2] zeigt, dass in der Zeit seit der Erstveröffentlichung des AFA 1980 pro Jahr durchschnittlich gut zwei Publikationen erschienen sind (siehe nachfolgende Tabelle). Ein Publikationspeak erreichte der AFA in den Jahren von 1988 bis 1991 mit knapp 20 Veröffentlichungen.

[2] PsychInfo und Web of Science, ergänzt mit der Literaturliste dieses Kapitels.

Tabelle 3.1
Anzahl der Publikationen zum AFA

Jahr	80 81	82 83	84 85	86 87	88 89	90 91	92 93	94 95	96 97	98 99	00 01	02 03	04 05	06 07	08 09
Anzahl Publikationen	2	3	3	5	9	8	5	4	4	2	3	6	1	5	5

Die Gründe für das geringe Interesse am AFA lassen sich an der von verschiedener Seite geäusserten Kritik an diesem Vorgehen aufzeigen. Die wichtigsten Beiträge dazu lieferten Angleitner und Demtröder (1988), Block (1989) und Moser (1989).

So sieht Block (1989) im AFA keineswegs „den neuen Weg, die Persönlichkeit zu erforschen" oder „den Lieferanten einer Vielzahl aufregender Forschungsprogramme" wie ihn Buss und Craik (1984) angepriesen haben, weil er kein geeignetes Verfahren ist, eine umfassende Beschreibung der Persönlichkeit zu liefern und vor allem keine Erklärung für konkretes Verhalten gibt. Moser (1989) stellt zudem in Frage, ob es ohne die Kenntnis von den dem Verhalten zugrunde liegenden Regeln möglich ist, von in vergangenen Situationen gezeigtem Verhalten auf zukünftiges Verhalten in anderen Situationen zu schliessen. Der AFA eignet sich gemäss Block (1989) auch nicht zur Beschreibung der internen Struktur einer Verhaltenskategorie, da die Beurteilerübereinstimmung beim Prototypenrating viel tiefer ist, als von Buss und Craik (1980) angegeben. Und auch als Validierungskriterium für bereits bestehende Persönlichkeitsskalen eignen sich mit dem AFA generierte Verhaltenskategorien – entgegen der Ankündigung von Buss und Craik (z. B. 1983b) – nur bedingt: Auf Grund ihrer Analysen gelangten Cooper et al. (1990) zur Erkenntnis, dass die psychometrische Qualität der Verhaltenskategorien anderen Skalen nicht so deutlich überlegen ist, als dass man sie uneingeschränkt als Vergleichsmassstab einsetzen könnte. Als grösste Einschränkung des AFA sehen Larsen und Buss (2001) jedoch dessen Theorielosigkeit: Es gibt keine Anleitung dazu, welche Verhaltensdimensionen wichtig sind, noch kann erklärt werden, weshalb sich Individuen in der Häufigkeit der konkret gezeigten Acts über die Zeit hinweg unterscheiden.

Nachfolgend führe ich ein paar weitere problematische Aspekte des AFA auf:

Laien als Quelle für das Ausgangsmaterial

Die Auskünfte von Laien als Ausgangsmaterial für die Beschreibung und Erklärung der menschlichen Persönlichkeit zu nehmen, wirft nach Moser (1989) die

Frage auf, ob das, was Laien als real existierend betrachten auch wirklich real existiert oder nur das Ergebnis eines sozialen Prozesses ist („*social psychological invasion*" der Persönlichkeit; Kenrick & Dantchik, 1983). Nach Block (1989) führt das alleinige Abstützen auf die Act-Nennungen der Laien auch dazu, dass bei einigen Acts der Zusammenhang zur Verhaltenskategorie nicht ersichtlich ist: Wieso sind zum Beispiel die von Buss und Craik aufgeführten Verhaltensweisen „Ich ass schnell ein Mittagessen" oder „Ich fuhr ein Motorrad" Beispiele für distanziertes Verhalten?

Unklarer Situationskontext

Ein häufig genannter Kritikpunkt am AFA betrifft die Unklarheit darüber, wie viel Kontextbeschreibung ein Act enthalten soll. Buss und Craik liessen den Kontext häufig offen, um die Situation nicht allzu sehr exklusiv werden zu lassen und so einer tiefen Basisrate entgegenzuwirken. Für Moser (1989) ist es jedoch nicht zulässig, den Situationskontext ausser Acht zu lassen, da es erst dieser erlaubt, Verhalten eindeutig Dispositionen zuzuordnen. Als Beispiel nennt er das Singen: Dies kann ein Zeichen der Freude sein – zum Beispiel nach einem Erfolg – oder aber auch ein Zeichen der Angst – Singen im dunklen Wald. Block (1989) fordert zudem auch die Erhebung der Motivation der Person, diese Handlung auszuführen, damit sich das Verhalten interpretieren lässt.

Larsen und Buss (2001) geben ein weiteres Beispiel, welches aufzeigt, wie wichtig die Kenntnis des genauen Situationskontextes ist, um ein Verhalten richtig interpretieren zu können: Sie gehen vom Act „Er insistierte darauf, dass ihn die anderen in sein Lieblings-Restaurant begleiteten." aus. Um diesen Act eindeutig der Dominanz-Kategorie zuordnen zu können, muss man Kenntnis zum Beispiel über die Beziehung der involvierten Personen untereinander, den Grund für das Essen oder die bisherigen Restaurantbesuche dieser Gruppe haben.

Block (1989) geht davon aus, dass die Probanden viele Acts nicht mit ja beantworten können, da die darin beschriebenen Situationen und das gezeigte Verhalten in dieser Kombination eher selten auftreten. Er nennt das Beispiel eines Acts, der beschreibt, dass das Treffen von seit langer Zeit nicht mehr gesehenen Freunden am Flughafen keine Emotionen auslöste. Die Wahrscheinlichkeit, dass eine Person in den vergangenen drei Monaten – ja überhaupt einmal im Leben – genau diese Situation erlebt hat, scheint sehr klein zu sein.

Das hier angesprochene Dilemma zwischen vager – und somit viele Personen ansprechender – und umfassender, genau definierter Situationsdarstellung lässt sich nicht auflösen. Buss und Craik entziehen sich der Auseinandersetzung

mit diesem Dilemma, indem sie explizit darauf hinweisen, dass es nicht ihr Ziel sei, Verhalten erklären zu wollen: „In field monitoring, this assessment approach sums displays of prototypical acts without regard to situational analysis or to attributions of causality to person, role, situation, or other factors. It remains strictly descriptive rather than explanatory" (Buss & Craik, 1989b, S. 399).

Multidimensionalität der Acts

In der Studie von Church et al. (2007) stuften die Probanden 69% der 198 Acts mehreren Verhaltenskategorien zu. Somit scheint das Phänomen der Überlappung von Verhaltenskategorien bei der Beschreibung der Persönlichkeit eher die Regel denn die Ausnahme zu sein. Diese auch in anderen Studien aufgetretene Multidimensionalität der Acts (z. B. Angleitner & Demtröder, 1988; Borkenau, 1986) stellt für Block (1989) ein ernsthaftes Problem des AFA dar. Er sieht die Gründe dafür in einer fehlerhaften Klassifikation oder in uneindeutigen Indikatoren der Verhaltenskategorien. Chruch et al. (2007) taxieren dies jedoch nicht als einen Fehler des AFA sondern als ein akkurates Abbild der Komplexität des Verhaltens, was sie anhand einiger Beispiele belegen. Einige Autoren (z. B. Angleitner & Demtröder, 1988; Buss & Craik, 1989) schlagen vor, dass sich multidimensional eingestufte Acts für die Erfassung mehrerer Dimensionen einsetzen lassen. Dies steht jedoch in einem Widerspruch zur Prämisse der traditionellen Persönlichkeitsdiagnostik, möglichst unabhängige Dimensionen zu erfassen.

Angleitner und Demtröder (1988) und Borkenau (1986) stellten entgegen ihrer Hypothese fest, dass Probanden Acts durchaus auch als hochprototypisch für zwei Kategorien einstuften. Da dies jedoch hauptsächlich bei semantisch ähnlichen Kategorien vorkommt, lässt sich dieses Phänomen mit der Hypothese der systematischen Überlappung von Borkenau (1986) erklären. Somit ist die Vorstellung der Zugehörigkeit zu einer einzelnen Kategorie wie sie Rosch (1975) für Objekte beschreibt, zumindest für semantisch nahe Verhaltenskategorien nicht zutreffend. Zudem weist Borkenau darauf hin, dass Aussagen von Buss und Craik wie „dispositional categories are composed of acts that differ in their within-category status from highly central or prototypical to progressively more peripheral until the fuzzy borders are reached and adjoining categories are entered" (Buss & Craik, 1985, S. 936) oder „at the borders, the array of peripheral acts for a given dispositional category blends into adjacent act categories" (Buss & Craik, 1983, S. 109) missverständlich sind, da sie andeuten, dass nur periphere Acts mehreren Dimensionen zugeordnet werden können.

Problem der retrospektiven Einschätzung des Verhaltens

Block (1989) führt an, dass bei der Befragung von Personen nach deren in der Vergangenheit gezeigtem Verhalten, Gedächtnisverzerrungen auftreten können, und kritisiert, dass Buss und Craik dies im AFA nirgends berücksichtigen. Die Zuverlässigkeit der subjektiven Einschätzung der Auftretenshäufigkeit bestimmter Verhaltensweisen war schon verschiedentlich Gegenstand von Forschungsarbeiten (z. B. Borkenau & Müller, 1992; Borkenau & Ostendorf, 1987; Kolar, Funder & Colvin, 1996). Gosling, John, Craik und Robins (1998) fanden in ihrer Metaanalyse über sechs Studien den eher bescheidenen Zusammenhang von ρ = .21 zwischen der Selbst- und der Fremdbeurteilung. Zudem konnten sie aufzeigen (siehe auch Borkenau & Ostendorf, 1987), dass die Akkuratesse dieser Einschätzungen – also die Übereinstimmung zwischen der Selbstauskunft und der Einstufung durch Beobachter – von der jeweiligen Verhaltenskategorie (so werden Verhaltensweisen der Kategorie Extraversion genauer eingeschätzt als solche der Verträglichkeit), deren Einfachheit der Beobachtbarkeit, der Auftretenshäufigkeit und der sozialen Erwünschtheit dieser Verhaltenskategorie und von der Persönlichkeit des Probanden abhängt. Amelang et al. (1989) wiesen zudem nach, dass der Zusammenhang zwischen Selbst- und Fremdbeurteilung bei hochprototypischen Acts deutlich höher ist, als bei niedrigprototypischen (r = .51 vs. r = .26).

Amelang et al. (1991) zeigten sich erstaunt über die zum Teil relativ hohe Basisrate einzelner Acts. Sie gehen davon aus, dass die Probanden bei der Bearbeitung der einzelnen Acts den angedeuteten Situationskontext eher generell als spezifisch wahrnehmen und/oder eher grosszügig sind, wenn es darum geht, das eigene Verhalten als identisch mit dem beschriebenen einzustufen. Diese Annahme unterstützen sie mit der Erkenntnis, dass der Grad der Spezifität der geschilderten Situation die Basisrate eines Acts beeinflusst. Zudem scheint bei der Einstufung die soziale Erwünschtheit der Acts eine bedeutende Rolle zu spielen: Die prototypischsten Acts sind eher diese, welche kaum jemand schon erlebt hat, welche jedoch auf grosse Zustimmung stossen (Block, 1989).

Für den Zweck der Itemgenerierung im Rahmen der Konstruktion eines Persönlichkeitstests sind jedoch die meisten oben aufgeführten Kritikpunkte nicht bedeutsam. In diesem Zusammenhang stellt sich die Frage, ob es möglich ist, Persönlichkeitsdimensionen mit dem AFA zu operationalisieren und wie gut die Übereinstimmung mit nach traditionellen Methoden entwickelten Persönlichkeitsskalen ist. Drauf gehe ich im nachfolgenden Kapitel ein.

3.6 Die Entwicklung von Persönlichkeitsskalen mit dem Act Frequency Approach

Die Methode des AFA setzten einige Forschergruppen für die Entwicklung von Persönlichkeitsskalen ein, da er eine systematische Methode zur Gerierung von Items bietet (Cooper et al., 1990), obwohl Buss und Craik (1989) dies nicht als dessen Ziel ansehen und sich gegen diesen Einsatzzweck aussprechen. Wie oben schon erwähnt, sehen sie es als besonders problematisch an, wenn Probanden in solchen Persönlichkeitsskalen auf einer hypothetischen Basis ihr Verhalten einstufen müssen, weil sie dieses so noch nie gezeigt haben oder noch nie zeigen konnten. Dabei sieht Block (1989) in der Bildung von „*multiple act criterions*" oder „*act trends*" nichts anderes als das, was Testentwickler schon bei einer herkömmlichen Konstruktion einer Persönlichkeitsskala durchführen. Was beim AFA jedoch fehlt, ist der langwierige Überarbeitungsprozess, welcher schlussendlich zu reliablen und validen Skalen führt. Zudem sieht er in einigen Acts keinen Unterschied zu traditionellen Persönlichkeits-Fragebogenitems: Inwiefern soll das Item „Ich sprach zu praktisch allen Leuten an der Party" deutlich spezifischer sein, als das Item „Ich tendiere dazu viel zu sprechen"? Zudem zeigte sich, dass der Zusammenhang zu herkömmlichen Persönlichkeitsinventaren um so höher ausfällt, je prototypischer eine Handlung für die betreffende Eigenschaft ist (Buss & Craik, 1980) und man sich deshalb die Frage stellen muss, ob sich der grosse Aufwand für die Erstellung eines Persönlichkeitstests mit dem AFA wirklich auch lohnt. Problematisch bei der Entwicklung von Persönlichkeits-Fragebogen mit dem AFA sind zudem die bei einigen Items auftretenden tiefen Basisraten und Mittelwerte, da durch den Prototypenansatz Persönlichkeitsmerkmale oft auch in Extremausprägungen charakterisiert werden (Hodapp, Gableske, Riedemann & Bongard, 2004; Krüger & Amelang, 1995). Block (1989) fordert deshalb – ganz im Sinne von Jaccard (1974) –, dass Acts, welche alle oder niemand mit ja beantworten, aus der Skala auszuschliessen sind.

Als besonders hilfreich und fruchtbar erwies sich der AFA bei der Operationalisierung von Konstrukten, welche noch über kein ausgereiftes theoretisches Grundgerüst verfügen oder allgemein als theoretisch schwer fassbar gelten (z. B. Amelang, Schwarz & Wegemund, 1989; Cooper et al., 1990; Krüger & Amelang, 1995; Romero, Luengo, Carrillo-de-la-Peña & Otero-López, 1994). Die in Acts kondensierten impliziten Theorien von Laien über eine Persönlichkeitseigenschaft verhelfen somit den Forschern trotz mangelhafter theoretischer Fundierung eine valide Persönlichkeitsskala zu entwickeln. So konnte Broughton (1984) aufzeigen, dass die Verwendung des Prototypenansatzes bei der Konstruktion von

Persönlichkeitsskalen zu einer höheren Validität führt, als sich dies mit anderen Ansätzen – unter anderem dem faktorenanalytischen Ansatz, der rationalen Testkonstruktion oder der Kontrastgruppen-Methode – erreichen lässt.

Nachfolgende Aufzählung vermittelt einen Eindruck von der Vielfalt der mit dem AFA entwickelten und in wissenschaftlichen Beiträgen referierten Persönlichkeits- und Einstellungsskalen:

- Distanziertheit, Geselligkeit, Dominanz und Unterwürfigkeit (Buss & Craik, 1981)
- Soziale Intelligenz (Amelang et al., 1989)
- Big Five (Botwin & Buss, 1989)
- Leistungsmotivation (Piedmont, 1989)
- Kreativität (Amelang et al., 1991)
- Hilflosigkeit (Peterson, 1993)
- Impulsivität, Extraversion und Neurotizismus (Romero et al., 1994)
- Risikobereitschaft (Krüger & Amelang, 1995)
- Rassismus („*anti-White attitudes*", Johnson & Lecci, 2003)
- Feindseligkeit (Hodapp et al., 2004)
- Dominanz und emotionale Affiliation (Muck et al., 2006)

Exemplarisch stelle ich nachfolgend das Vorgehen zur Entwicklung eines Persönlichkeits-Fragebogens mit Hilfe des AFA am Beispiel des Risikobereitschafts-Fragebogens von Krüger und Amelang (1995) dar, welche sich streng an das von Buss und Craik (1981) beschriebene Vorgehen hielten.

Die Autoren wählten mit der Risikobereitschaft ein facettenreiches Konstrukt, dessen Existenz offensichtlich erscheint und das zur Erklärung bestimmter Verhaltensweisen auch zweckmässig ist, zu welchem aber bislang nur die Entwicklung von Instrumenten mit mässig hohen Validitäten gelang, was an der schwierigen theoretischen Eingrenzung und Erfassung des Konstruktes liegt. So konnten die auf diesem Gebiet tätigen Forscher auch noch nicht den Beweis erbringen, dass es sich bei der Risikobereitschaft um ein homogenes, stabiles und konsistentes Persönlichkeitsmerkmal handelt.

Die eine Skalenkonstruktion besonders erschwerende Theorielosigkeit umgingen Krüger und Amelang, indem sie mit dem AFA auf die impliziten Vorstellungen von psychologischen Laien zurückgriffen und somit keine a priori festgelegte Definition von Risikobereitschaft benötigten. Für die Generierung der Acts

rekrutierten die Autoren nach dem Schneeballprinzip 38 Personen im Alter von 16 bis 71 Jahren. In der Instruktion erhielten sie verschiedene Definitionen von Risikobereitschaft, zum Beispiel „Bereitschaft, sich in potenziell gefährliche Situationen zu begeben" oder „unverzügliches Fällen von Entscheidungen trotz Ungewissheit". Die Versuchspersonen wurden gebeten, sich

> mindestens je eine weibliche und eine männliche Personen aus dem Freundes- und Bekanntenkreis vorzustellen, die sie als risikobereit einschätzten. Für jede dieser Personen sollten sie drei einzelne Handlungen in konkreten Situationen schildern, in denen ihrer Meinung nach das riskante Verhalten dieser Person zum Ausdruck gekommen ist. (Krüger & Amelang, 1995, S. 38)

Weiter gaben die Autoren Beispiele für erwünschte Verhaltensbeschreibungen in konkreten Situationen (z. B. Er/Sie verteilte die Rollen für ein Theaterstück.) und für die unerwünschten Verhaltenszuschreibungen (z. B. Er/Sie hat ein herrschsüchtiges Wesen.) anhand des Eigenschaftskonstruktes Dominanz.

Aus den 323 auf diese Weise generierten Acts wählten die Forscher diejenigen aus, welche folgende Kriterien erfüllten:

1. Sie bestehen nicht aus allgemeinen Trendaussagen (z. B. „Er ist ein mutiger Mensch"), sondern beschreiben konkrete und prinzipiell beobachtbare Verhaltensweisen („Er kletterte auf einem ungesicherten Baugerüst herum").
2. Sie enthalten zumindest indirekt einen Situationskontext („Er wich einer ernsthaften Aussprache mit Schmeicheleien aus").
3. Sie können prinzipiell von jedermann gezeigt werden und sind nicht an Alter, Geschlecht, Beruf oder soziale Stellung gebunden (wie etwa „Sie liess sich an der Gebärmutter operieren").

Einzelne Aussagen modifizierten die Autoren soweit, dass sie diesen Kriterien genügten, ohne jedoch den inhaltlichen Sinn der jeweiligen Aussage zu verändern. Zudem korrigierten sie Fehler in der Rechtschreibung, Zeichensetzung und im Satzbau, kürzten lange Aussagen, brachten alle Nennungen ins Perfekt und setzten statt Namen „er" oder „sie" ein. Bei ähnlichen Verhaltensbeispielen wählten sie nur eines. Auf diese Weise reduzierten sie die Liste auf 200 Acts und legten sie in zufälliger Reihenfolge weiteren 39 Probanden mit folgender Anweisung zur Einstufung der Prototypizität vor:

> Auf den folgenden Seiten finden Sie eine Liste von konkreten Handlungen, die mehr oder weniger gut die Eigenschaft „Risikobereitschaft" beschrei-

> ben. Wir sind daran interessiert, wie gut diese Verhaltensbeispiele Ihrer Meinung nach Risikobereitschaft repräsentieren, für wie geeignet Sie also jede dieser Handlungen zur Charakterisierung risikoreicher Menschen halten. (Krüger & Amelang, 1995, S. 40)

Die Anweisung enthielt zusätzlich eine Umschreibung des Konstruktes Risikobereitschaft und Beispiele zum Gebrauch der Einschätzungsskala, welche von 7 = „sehr prototypisch" bis 1 = „gar nicht prototypisch" reicht.

Um für die nachfolgenden Berechnungen eine möglichst gute Differenzierung zwischen hoch- und niedrigprototypischen Acts zu gewährleisten, wählten die Autoren aus den 200 nach Prototypizität rangierten Acts die 28 hoch- (M = 5.5, SD = 1.0) und 19 niedrigprototypischsten (M = 2.6, SD = 0.8) aus, indem sie in einem ersten Schritt 111 Acts mit einem Prototypenrating zwischen 3.5 und 4.9 ausschlossen und in einem zweiten Schritt weitere 42 Items, bei denen die Beurteilungen am wenigsten übereinstimmten, deren Prototypizität mit Merkmalen der Beurteiler (Geschlecht oder Experte/Laie) und/oder mit dem Geschlecht der im Act handelnden Person zusammenhing. In Tabelle 3.2 sind beispielhaft je vier hoch- und niedrigprototypische Acts aufgeführt.

Tabelle 3.2

Hoch- und niedrigprototypische Acts zur Eigenschaft „Risikobereitschaft" (Krüger & Amelang, 1995, S. 51-52)

	Prototypizität *M* (*SD*)	Basisrate (%)
Ich sagte mich von der Gruppe los und ging allein durch die Wüste.	6.2 (1.3)	5
Um in den Abgrund schauen zu können, ging ich bis zum Ende des Felsvorsprunges, ohne zu wissen, ob dieser überhaupt befestigt war.	6.0 (1.6)	32
Kurz vor einer engen Kurve überholte ich noch ein anderes Fahrzeug.	5.8 (2.0)	43
Ich baute mich vor einer Gruppe von Schlägern auf.	5.3 (1.9)	14
Als mir jemand seine Gefühle mitteilte, lachte ich ihn aus.	2.5 (1.7)	22
Vor meinem Chef bagatellisierte ich die Probleme der Mitarbeiterinnen.	2.3 (1.5)	37
Als das Telefon klingelte, liess ich mich verleugnen.	1.8 (1.3)	81
Als die anderen gelobt wurden, warf ich ihnen missgünstige Blicke zu.	1.5 (0.9)	66

Für die Entwicklung der definitiven Version des actbasierten Fragebogens zu Risikobereitschaft führten Krüger und Amelang auf der Basis der erhobenen Daten zusätzliche Berechnungen zur Itemselektion durch: Ein Kriterium, welches schon Jaccard (1974) erwähnt hatte, betrifft die Basisrate des Acts, das heisst den Anteil an Probanden, welche angegeben hat, die beschriebene Handlung in der Vergangenheit mindestens einmal ausgeführt zu haben. Krüger und Amelang setzten als Selektions-Grenze einen Wert von 5%, den elf Acts nicht erreichten. So strichen sie zum Beispiel den Act „Im Dschungel übernachtete ich unter freiem Himmel", welcher eine Basisrate von lediglich 3% erzielte. Als weiteres Kriterium zogen die Autoren den Grad der Übereinstimmung der Act-Ratings von Fremdbeurteilern hinzu: Jeder Studienteilnehmer hatte drei ihm bekannte Personen anzugeben, welche diesen anhand desselben Fragebogens einschätzten. Als Mass für die Übereinstimmung der Fremdbeurteilungsratings verwendeten die Autoren Cronbachs Alpha und setzten den Grenzwert bei .20. Fünf Acts schafften diese Hürde nicht. Die definitive Version des Fragebogens zur Erfassung von Risikobereitschaft bestand aus 15 hoch- und 16 niedrigprototypischen Acts, welche die Autoren getrennt und ungewichtet zu Skalen zusammenfassten. Die Homogenität der beiden Skalen beträgt α = .55 respektive .54 und die einzelnen Acts korrelieren im Schnitt mit r = .12, beides Anzeichen dafür, dass das Risikokonstrukt mit topografisch sehr unterschiedlichen Handlungen operationalisiert wurde, was es potenziell ermöglicht, mit dieser Skala differenziertes Verhalten zu erfassen. Buss und Craik (1983a) sprechen in diesem Zusammenhang von einer lockeren Struktur des fraglichen Konstruktes. Zusammen mit weiteren Verfahren zur Einschätzung von Risikobereitschaft bearbeiteten 101 Probanden die beiden Skalen. Dabei mussten sie angeben, ob sie das geschilderte Verhalten „noch nie", „selten", „mehrmals" oder „oft" selbst ausgeführt hatten (Buss & Craik, 1980).

Auch wenn sich der AFA nicht – wie von Buss und Craik proklamiert – als das Wundermittel zur Erforschung der Persönlichkeit herausgestellt hat, zeigen oben aufgeführte Beispiele, dass er bei der Messung von theoretisch schwer fassbaren Persönlichkeitskonstrukten ein wertvolles Hilfsmittel darstellt. Und auch wenn der AFA sich nicht zu den Methoden der rationalen Testkonstruktion im engeren Sinne zählen lässt, hilft er diesen Prozess zu systematisieren und befreit den Testentwickler weitgehend von der schwierigen Suche nach den richtigen Itemformulierungen, indem dies die Probanden – also quasi *Subject Matter Experts* – übernehmen. Damit ist der AFA bestens für den Einsatz im Rahmen der bewerberzentrierten Psychodiagnostik (Boss, 2005) geeignet, indem sich damit Items generieren lassen, die sehr nahe an der Erfahrungsrealität der zu untersuchenden Population liegen. So ist es auch grundsätzlich möglich, für

unterschiedliche Gruppen massgeschneiderte Tests für die Erfassung desselben Konstruktes zu erstellen. Auf Grund dieser Überlegungen ist nicht ganz nachvollziehbar, dass Forscher den AFA nicht häufiger für die Entwicklung von Persönlichkeitsskalen einsetzen. So wählten zum Beispiel auch die Autoren des Bochumer Inventars zur berufsbezogenen Persönlichkeitsbeschreibung (BIP; Hossiep & Paschen, 1998) bei der Generierung der Items die traditionelle Vorgehensweise, indem sie diese auf Grund der inhaltlichen Definitionen der gewählten Persönlichkeitsdimensionen quasi im „stillen Kämmerchen" formulierten. Dies stellt zwar ein sehr effizientes Vorgehen dar, bringt jedoch den Nachteil mit sich, dass die Items eher althergebracht und beliebig ausfallen, so dass beim späteren Fragebogenbearbeiter rasch ein Déjà-vu-Eindruck und nach einer gewissen Zeit Langeweile entsteht. So ist zum Beispiel das Item „Zuweilen verhalte ich mich sehr dominant gegenüber anderen." (nach Hossiep & Paschen, 1998) doch deutlich unspektakulärer als „Bei der Teambesprechung wies sie die anderen darauf hin, dass sie so viele Abschlüsse tätigen würde wie alle anderen zusammen." (Muck et al., 2002, S. 83).

3.7 Literaturverzeichnis

Allen, G. (1993). An application of the act frequency approach in the study of person–job fit. *Library & Information Science Research, 15*, 249–255.

Allport, G. W. (1937). *Personality: A psychological interpretation*. New York, NY: Holt, Rinehart and Winston.

Alston, W. P. (1975). Traits, consistency and conceptual alternatives for personality theory. *Journal for the Theory of Social Behaviour, 5*, 17–48.

Amelang, M. & Bartussek, D. (1990). *Differentielle Psychologie und Persönlichkeitsforschung* (3. Auflg.). Stuttgart: Kohlhammer.

Amelang, M., Bartussek, D., Stemmler, G. & Hagemann, D. (2006). *Differentielle Psychologie und Persönlichkeitsforschung* (6., vollst. überarb. Auflg.). Stuttgart: Kohlhammer.

Amelang, M., Herboth, G., & Oefner, I. (1991). A prototype strategy for the construction of a creativity scale. *European Journal of Personality, 5*, 261–285.

Amelang, M. & Schmidt-Atzert, L. (2006). *Psychologische Diagnostik und Intervention.* Berlin: Springer.

Amelang, M., Schwarz, G. & Wegemund, A. (1989). Soziale Intelligenz als Trait-Konstrukt und Test-Konzept bei der Analyse von Verhaltenshäufigkeiten. *Zeitschrift für Differentielle und Diagnostische Psychologie, 10*, 37–57.

Angleitner, A., Buss, D. A., & Demtröder, A. I. (1990). A cross-cultural comparison using the Act Frequency Approach (AFA) in West Germany and the United States. *European Journal of Personality, 4*, 187–207.

Angleitner, A., & Demtröder, A. I. (1988). Acts and dispositions: A reconsideration of the Act Frequency Approach. *European Journal of Personality, 2*, 121–141.

Asendorpf, J. B. (1999). *Psychologie der Persönlichkeit* (2. Auflg.). Berlin: Springer.

Bandura, A. & Walters, R. H. (1963). *Social learning and personality development*. New York, NY: Holt, Rinehart and Winston.

Block, J. (1989). Critique of the Act Frequency Approach to personality. *Journal of Personality and Social Psychology, 56*, 234–245.

Borkenau, P. (1986). Toward an understanding of trait intercorrelations: Acts as instances for several traits. *Journal of Personality and Social Psychology, 51*, 371–381.

Borkenau, P., & Müller, B. (1992). Inferring act frequencies and traits from behavioral observations. *Journal of Personality, 60*, 553–573.

Borkenau, P., & Ostendorf, E. (1987). Retrospective estimates of act frequencies: How accurately do they reflect reality? *Journal of Personality and Social Psychology, 52*, 626–638.

Boss, P. (2005). Assessment in der Arbeitswelt – Kriterien für eine bewerberzentrierte Personalauswahl. In M. Rehbinder (Hrsg.), *Psychologische Aspekte im Recht der Personalführung* (S. 21–45). Bern: Stämpfli.

Botwin, M. D., & Buss, D. M. (1989). Structure of act-report data: Is the five-factor model of personality recaptured? *Journal of Personality and Social Psychology, 56,* 988–1001.

Broughton, R. (1984). A prototype strategy for construction of personality scales. *Journal of Personality and Social Psychology, 47*, 1334–1346.

Buss, D. M. (1988). The evolution of human intrasexual competition: Tactics of mate attraction. *Journal of Personality and Social Psychology, 54*, 616–628.

Buss, D. M., & Craik, K. H. (1980). The frequency concept of disposition: Dominance and prototypically dominant acts. *Journal of Personality, 48,* 379–392.

Buss, D. M., & Craik, K. H. (1981). The act frequency analysis of interpersonal dispositions: Aloofness, gregariousness, dominance and submissiveness. *Journal of Personality, 49,* 175–192.

Buss, D. M., & Craik, K. H. (1983a). The Act Frequency Approach to personality. *Psychological Review, 90,* 105–126.

Buss, D. M., & Craik, K. H. (1983b). The dispositional analysis of everyday conduct. *Journal of Personality, 51,* 393–412.

Buss, D. M., & Craik, K. H. (1983c). Act prediction and the conceptual analysis of personality scales: Indices of act density, bipolarity, and extensity. *Journal of Personality and Social Psychology, 45,* 1081–1095.

Buss, D. M., & Craik, K. H. (1984). Acts, dispositions, and personality. In B. A. Maher & W. B. Maher (Eds.), *Progress in experimental personality research* (Vol. 13, pp. 242–301). New York, NY: Academic Press.

Buss, D. M., & Craik, K. H. (1985). Why *not* measure that trait? Alternative criteria for identifying important dispositions. *Journal of Personality and Social Psychology, 48,* 934–946.

Buss, D. M., & Craik, K. H. (1986a). The Act Frequency Approach and the construction of personality. In A. Angleitner, A. Furnham, & G. Van Heck (Eds.), *Personality psychology in Europe. Volume 2. Current trends and controversies* (pp. 141–156). New York, NY: Academic Press.

Buss, D. M., & Craik, K. H. (1986b). Acts, dispositions, and clinical assessment: The psychopathology of everyday conduct. *Clinical Psychology Review, 6,* 387–406.

Buss, D. M., & Craik, K. H. (1989). On the cross-cultural examination of acts and dispositions. *European Journal of Personality, 3*, 19–30.

Cantor, N., & Mischel, W. (1977). Traits and prototypes: Effects on recognition memory. *Journal of Personality and Social Psychology, 35,* 38–48.

Church, A. T., Katigbak, M. S., Miramontes, L. G., Del Prado, A. M., & Cabrera, H. F. (2007). Culture and the behavioural manifestations of traits: An application of the Act Frequency Approach. *European Journal of Psychology, 21*, 389–417.

Cinite, I., Duxbury, L. E., & Higgins, C. (2009). Measurement of perceived organ-

isational readiness for change in the public sector. *British Journal of Management, 20*, 265–277.

Cooper, W. H., Dyke, l., & Kay, P. (1990). Developing act frequency measures of organizational behaviors. In L. R. Jauch & J. L. Wall (Eds.), *Academy of Management Best Papers Proceedings* (pp. 396–399). Ada, OH: The *Academy of Management.*

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Endler, N. S., Hunt, J. McV., & Rosenstein, A. J. (1962). An S-R inventory of anxiousness. *Psychological Monographs, 76*, 1–33.

Fishbein, M. (1972). The prediction of behaviors from attitudinal variables. In K. K. Sereno & C. D. Mortensen (Eds.), *Advances in communication research* (pp. 3–31). New York, NY: Harper and Row.

Gosling, S. D., John, O. P., Craik, K. H., & Robins, R. W. (1998). Do people know how they behave? Self-reported act frequencies compared with on-line codings by observer. *Journal of Personality and Social Psychology, 74*, 1337–1349.

Gough, H. G. (1957). *California Psychological Inventory manual*. Palo Alto, CA: Consulting Psychologists Press.

Hampshire, S. (1953). Dispositions. *Analysis, 14*, 5–11.

Hodapp, V., Gableske, K., Riedemann, P. & Bongard, S. (2004). Konstruktion und Validierung eines Feindseligkeitsfragebogens auf der Basis des Handlungs-Häufigkeits-Ansatzes. *Zeitschrift für Klinische Psychologie, Psychiatrie und Psychotherapie, 52,* 346–358.

Höft, S. (2002). *Grundlagen einer persönlichkeitsorientierten Berufseignungsdiagnostik: verhaltens- und berufsbezogene Aspekte des Fünf-Faktoren-Modells der Persönlichkeit.* Berlin: Dissertation.de

Hossiep, R. & Paschen, M. (1998). *Das Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung*. Göttingen: Hogrefe.

Jaccard, J. J. (1974). Predicting social behavior from personality traits. *Journal of Research in Personality, 7*, 358–367.

Jackson, D. N. (1971). The dynamics of structured personality tests: 1971. *Psychological Review, 78*, 229–248.

Johnson, J. D., & Lecci, L. (2003). Assessing anti-White attitudes and predictiong percieved racism: The Johnson-Lecci Scale. *Personality and Social Psy-*

chology Bulletin, 29, 299–312.

Kenrick, D. T., & Dantchik, A. (1983). Interactionism, idiographics, and the social psychological invasion of personality. *Journal of Personality, 51*, 286–307.

Kolar, D. W., Funder, D. C., & Colvin, C. R. (1996). Comparing the accuracy of personality judgments by the self and knowledgeable others. *Journal of Personality, 64*, 311–337.

Krüger, C. & Amelang, M. (1995). Bereitschaft zu riskantem Verhalten als Trait-Konstrukt und Test-Konzept: Zur Entwicklung eines Fragebogens auf der Basis des Handlungs-Häufigkeits-Ansatzes. *Diagnostica, 41*, 35–52.

Larsen, R. J., & Buss, D. M. (2001). *Personality psychology. Domains of knowledge about human nature.* Boston, MA: McGraw-Hill.

Lorge, I. (1935). Personality traits by fiat. I. The analysis of the total trait scores and keys of the Bernreuter Personality Inventory. *Journal of Educational Psychology, 26*, 273–278.

Magnusson, D., & Endler, N. S. (1977). Interactional psychology: Present status and future prospects. In D. Magnusson & N. S. Endler (Hrsg.), *Personality at the crossroads* (pp. 3–36). Hillsdale, NJ: Erlbaum.

Mischel, W. (1968). *Personality and assessment*. New York, NY: Wiley.

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review, 80*, 252–283.

Mischel, W., & Peake, P. K. (1982). Beyond déjà vu in the search for cross-situational consistency. *Psychological Review, 89*, 730–755.

Moser, K. (1989). The Act-Frequency Approach: *A conceptual critique. Personality and Social Psychology Bulletin, 15*, 73–83.

Muck, P. M., Höft, S., Hell, B. & Schuler, H. (2006). Die Konstruktion eines berufsbezogenen Persönlichkeitsfragebogens. Integration von Interpersonalem Circumplex, Fünf-Faktoren-Modell und Act Frequency Approach. *Diagnostica, 52*, 76–87.

Pervin, L. A., & John, O. P. (2001). *Personality. Theory and research* (8th ed.). New York, NY: Wiley.

Peterson, C. (1993). Helpless behaviour. *Behaviour Research and Therapy, 31,* 289–295.

Piedmont, R. L. (1989). The Life Activities Achievement Scale: An act-frequency

approach to the measurement of motivation. *Educational and Psychological Measurement, 49*, 863–874.

Romero, E., Luengo, M. A., Carrillo-de-la-Peña, M. T., & Otero-López, J. M. (1994). The Act Frequency Approach to the study of impulsivity. *European Journal of Personality, 8*, 119–133.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology, 7*, 532–547.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Loyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573–605.

Ryle, G. (1949). *The concept of mind.* New York, NY: Barnes & Noble.

Shopshire, M. S., & Craik, K. H. (1996). An act-based conceptual analysis of the obsessive-compulsive, paranoid, and histrionic personality disorders. *Journal of Personality Disorders, 10*, 203–218.

Smid, N., Douma, M., Van Lenthe, J., & Ranchor, A. (1988). The predictive validity of three different types of personality assessment instruments. *European Journal of Personality, 2*, 143–154.

Sprock, J. (2000). Gender-typed behavioral examples of histrionic personality disorder. *Journal of Psychopathology and Behavioral Assessment, 22*, 107–122.

Szamosi, L. T., & Duxbury, L. (2002). Development of a measure to assess organizational change*. Journal of Organizational Change Management, 15*, 184–201.

Wallace, J. (1966). An abilities conception of personality: Some implications for personality measurement. *American Psychologist, 21*, 132–138.

Wiggins, J. S. (1979). A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of Personality and Social Psychology, 37*, 395–412.

Willmann, E., Feldt, K., & Amelang, M. (1997). Prototypical behaviour patterns of social intelligence: An intercultural comparison between Chinese and German subjects. *International Journal of Psychology, 32*, 329–346.

Zadeh, L. A., Fu, K. S., Tanaka, K., & Shimura, M. (Eds.). (1975). *Fuzzy sets and their applications to cognitive and decision processes.* New York, NY: Academic Press.

4. Das Wertequadrat

4.1 Die Kernaussagen des Wertequadrates

Als zweites Konstruktionsprinzip setzte unser Entwicklungsteam beim Leadership-Fragebogen das Wertequadrat (Helwig, 1948) ein, welches uns als Raster für die Formulierung verschiedener Verhaltensmöglichkeiten in den mit dem Act Frequency Approach gesammelten Situationen diente.

Das Persönlichkeitsmodell, welches dem Act Frequency Approach zugrunde liegt, unterstreicht die Wichtigkeit der Situation als Moderator menschlichen Verhaltens, indem die Auftretenshäufigkeit einer Persönlichkeitseigenschaft in verschiedenen Situationen ein Mass für deren Ausprägung darstellt. Einer Person wird demnach die Eigenschaft Spontaneität zugeschrieben, wenn sie am Arbeitsplatz, im Sportclub, in der Familie und in weiteren sozialen Situationen häufig spontan reagiert. Von der Alltagserfahrung ausgehend, nehmen wir jedoch von einer als spontan beschriebenen Person nicht an, dass es ihrem Naturell entsprechen würde, auf Grund eines Spontanentscheides ihren Arbeitsplatz und ihre Familie zu verlassen, um im Ausland eine neue Existenz aufzubauen. Hier würden wir wohl trotz allgemein zugeschriebener Spontaneität eine gewisse Besonnenheit erwarten, indem diese Person die Vor- und Nachteile einer solchen Aktion gründlich gegeneinander abwägt.

Dieses Beispiel zeigt auf, dass sich innerhalb der von der Gesellschaft definierten sozialen Norm liegendes Verhalten nicht per se den Kategorien „gut" oder „schlecht" zuordnen lässt, sondern situationsbezogen als „zweckmässig, angebracht" oder „unzweckmässig, unangebracht" einzustufen ist. Ein Modell, in welchem diese Überlegung umgesetzt ist und Verhalten nicht in der Polarität gut – schlecht eingeordnet wird, ist das Wertequadrat von Helwig (1936, 1948, 1951, 1967). Hier stehen zwei sich wechselseitig ergänzende (komplementäre) Werte oder Persönlichkeitseigenschaften in einer Balance respektive in einer dialektischen Ergänzung gegenüber. Diese Balance ist charakterisiert durch ein Spannungsverhältnis, in welchem die beiden untrennbar zueinander gehörenden Eigenschaften stehen. Bezogen auf unser Beispiel könnten wir diese Ergänzung der Spontaneität als Besonnenheit bezeichnen. Nach diesem Eigenschaftsverständnis lässt sich eine Person nicht generell als spontan oder als besonnen bezeichnen, sondern ihr Verhalten pendelt zwischen diesen beiden Verhaltensweisen, wobei es von der jeweiligen Situation abhängt, ob sie sich eher spontan oder eher besonnen verhält. Ist jedoch ein persönliches Merkmal oder ein Wert

im Verhaltensrepertoire einer Person nur ganz schwach oder gar nicht ausgebildet, so kippt der entsprechende Gegenwert auf Grund der fehlenden Ergänzung, des fehlenden Gegengewichtes in eine negative Ausprägung, die Übertreibung, wie dies im oben aufgeführten Beispiel der Fall ist: So würden wir eine Person, welche ohne vorausgehende Anzeichen von heute auf morgen ihren Job und ihre Familie verlässt, nicht mehr als spontan sondern eher als impulsiv oder unkontrolliert bezeichnen und damit das Verhalten als gesellschaftlich unakzeptiert etikettieren. Die Übertreibung der Eigenschaft Besonnenheit könnten wir als Über-Bedächtigkeit bezeichnen, womit alle vier Ausprägungen – ich bezeichne sie als Wertequadranten – des Wertequadrates definiert wären.

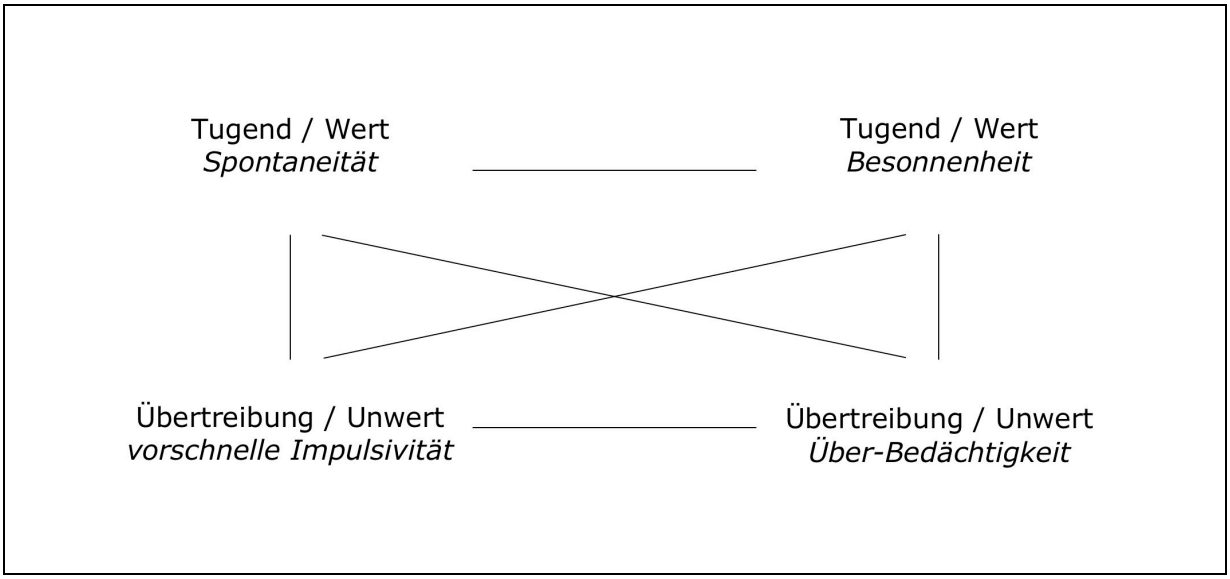


Abbildung 4.1 Der Aufbau des Wertequadrates (nach Schulz von Thun, 1989 und Gloor, 1993).

Der Hauptnutzen des Wertequadrates liegt vor allem darin, dass es die Beziehung zwischen im Alltagsverständnis entgegengesetzten Verhaltensweisen oder Persönlichkeitseigenschaften aufzeigt: Im in Abbildung 4.1 dargestellten Wertequadrat bilden die Persönlichkeitseigenschaften Spontaneität und Besonnenheit Tugenden, respektive sozial erwünschte Eigenschaften. Dabei stellt Spontaneität den positiven Gegenpol zur Eigenschaft oder zum Unwert Über-Bedächtigkeit dar, welche als Übertreibung der Besonnenheit angesehen werden kann. Die Übertreibung von Spontaneität ist die vorschnelle Impulsivität, welche ihrerseits den negativen Gegenpol zur Besonnenheit darstellt. Somit ist jede der beiden erwünschten Eigenschaften doppelt gegensätzlich präzisiert: Die Spontaneität

lässt sich gegen ihre Übertreibung, die vorschnelle Impulsivität, wie auch gegen die entgegengesetzte Übertreibung, die Über-Bedächtigkeit, abgrenzen.

Schulz von Thun (1989) erkannte das Potenzial dieses Denkansatzes für das Verständnis und die Analyse menschlichen Verhaltens. Bei Interventionen lässt er sich als Werkzeug einsetzen, um „die Entwicklungsrichtungen eines Menschen (oder auch einer Gruppe) zu bestimmen, die angezeigt sind, um den besonderen Herausforderungen der jeweiligen Berufspraxis und Lebenswelt gerecht zu werden" (Schulz von Thun, 1989, S. 47). Das grundsätzlich Neue an diesem Ansatz sah er darin, dass es bei der Personalentwicklung nun nicht mehr darum geht, Menschen von einem „schlechten" Verhalten zu einem „guten" zu führen, „sondern von dem Guten, wovon sie zuviel haben, hin zu dem Guten, welches ergänzend dazukommen müsste und vielleicht noch unterentwickelt ist" (Schulz von Thun, 2000, S. 54). Einen ähnlichen Entwicklungsansatz hat Groeben schon 1981 (siehe auch Erb, 1992) mit seinem Konzept der polaren Integration formuliert. Er überwindet dabei die bei gegenläufig auftretenden Persönlichkeitszügen üblicherweise vorgenommene Dichotomisierung, indem als Entwicklungsziel beide Pole einen möglichst starken Ausprägungsgrad erreichen sollen.

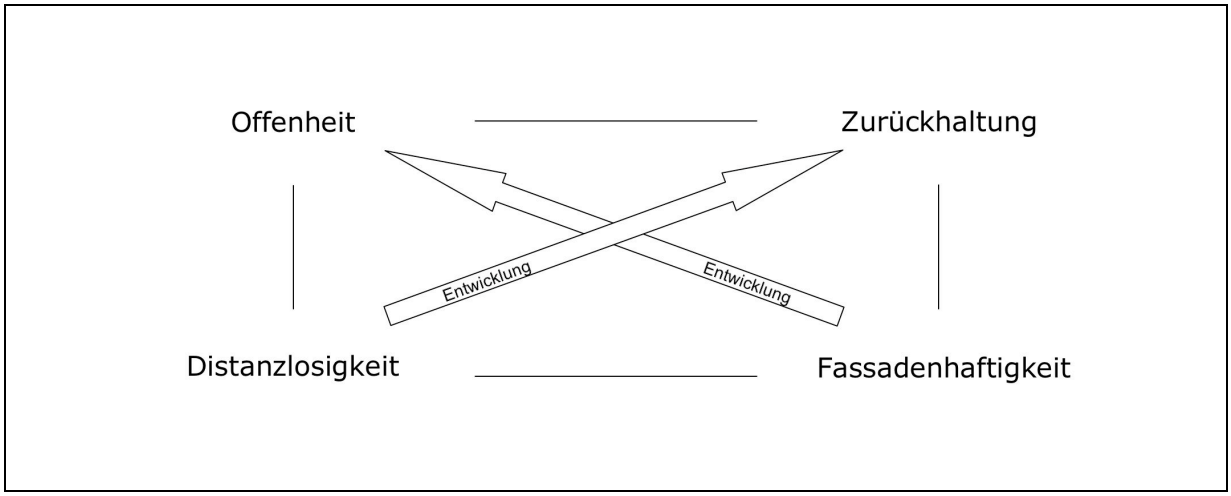


Abbildung 4.2 Das Wertequadrat als Entwicklungsquadrat (nach Gloor, 2007b, S. 111).

Abbildung 4.2 zeigt den Einsatz des Wertequadrates bei psychologischen Interventionen: Helwig (1948) nimmt an, dass es dann zu einem „entarteten" Verhalten kommt, wenn die beiden Tugenden nicht mehr in einer ausgeglichenen Balance zueinander stehen, weil eine Tugend nicht genügend entwickelt ist und somit die von ihm so genannte „Gegenspannung" (S. 123), das polare Span-

nungsverhältnis durch eine der Tugenden fehlt. Dies führt dazu, dass eine Verhaltenstendenz, zum Beispiel die Offenheit, ohne ihren Gegenspieler, die Zurückhaltung, übertrieben zum Ausdruck kommt und „vertikal abgleitet" (S. 123) in die Distanzlosigkeit. Um die Balance wiederherzustellen, muss nach diesem Denkansatz die der Übertreibung diagonal gegenüberliegende Tugend, hier die Zurückhaltung, entwickelt werden, wobei die positive Eigenschaft der übertriebenen Tugend nicht verloren gehen darf, da die betreffende Person ansonsten von einem Extrem ins andere kippt, also von der Distanzlosigkeit in die Fassadenhaftigkeit, und es zu einer Überkompensation des Verhaltens kommt. Umgekehrt ist es so, dass ein Wert nur gesteigert werden kann, wenn der positive Gegenwert sich entsprechend mitentwickelt. Somit kann sich jede Stärke, Kompetenz, Tugend je nach Situation zu einer Schwäche wandeln, wenn sie übertrieben wird. Anhand dieser Überlegungen formulierte Helwig (1948, S.123) das Wertegesetz: „Jeder Wert ist nur in ausgehaltener Spannung zu seinem positiven Gegenwert ein wirklicher Wert. ... Kein Wert ist an sich allein schon, was er sein soll – er wird es erst durch Einbeziehung des positiven Gegenwertes." Nach Schulz von Thun (1989) ist der Idealzustand eine dynamische Balance zwischen den beiden positiven Gegenwerten, so dass einer Person beide Kategorien von Verhaltensweisen zur Verfügung stehen, um sich auch in unterschiedlichsten Situationen adäquat verhalten zu können.

Beinahe zeitgleich mit Schulz von Thun beschrieb Ofman (1992) das Modell der Kernqualitäten in seinem Buch „Bezieling en kwaliteit in organisaties" (2005 in deutscher Sprache unter dem Titel „Qualität und Inspiration, Zugangswege zur Kreativität" erschienen), welches in den Niederlanden heute als ein Klassiker der Managementliteratur gilt. Das Modell liefert einen Ansatzpunkt für diverse Beratungs- und Abklärungssituationen wie zum Beispiel Coaching, Konfliktmanagement, Selektion, Teamentwicklung oder 360-Grad-Feedback, indem es dazu dient, die Dynamik in der Ausprägung von Persönlichkeitseigenschaften bildhaft aufzuzeigen.

Kernqualitäten beschreiben die charakteristischen positiven Eigenschaften oder Stärken eines einzelnen Menschen, wie zum Beispiel Tatkraft, Fürsorglichkeit, Ordentlichkeit oder Einfühlungsvermögen, die als angeboren und entwicklungsfähig angesehen werden. Die Übertreibung im Wertequadrat bezeichnet Ofman als Falle oder Schwachpunkt, eine über das Ziel hinausgeschossene Kernqualität: Flexibilität kann sich so in Launenhaftigkeit verformen. Die positiv gegenübergestellte Qualität der Falle nennt Ofman Herausforderung, in unserem Beispiel wäre dies Standfestigkeit. Kernqualität und Herausforderung sind einander ergänzende Qualitäten, welche idealerweise in einer Balance stehen. Fehlt diese Balance, so verformt sich unter Umständen die Kernqualität in die Falle.

Weiter postuliert Ofman, dass Menschen negativ auf ein Zuviel ihrer Herausforderung reagieren. Die Übertreibung der Herausforderung bezeichnet er deshalb als Allergie, den vierten Quadranten des Kernquadrates (siehe Abbildung 4.3). Die Allergie zu Flexibilität wäre demnach Starrheit. Für Ofman ist das Ziel einer Persönlichkeitsentwicklung die Herstellung einer ausgewogenen Balance zwischen Kernqualität und Herausforderung, damit Allergie und Falle möglichst schwach ausgeprägt sind.

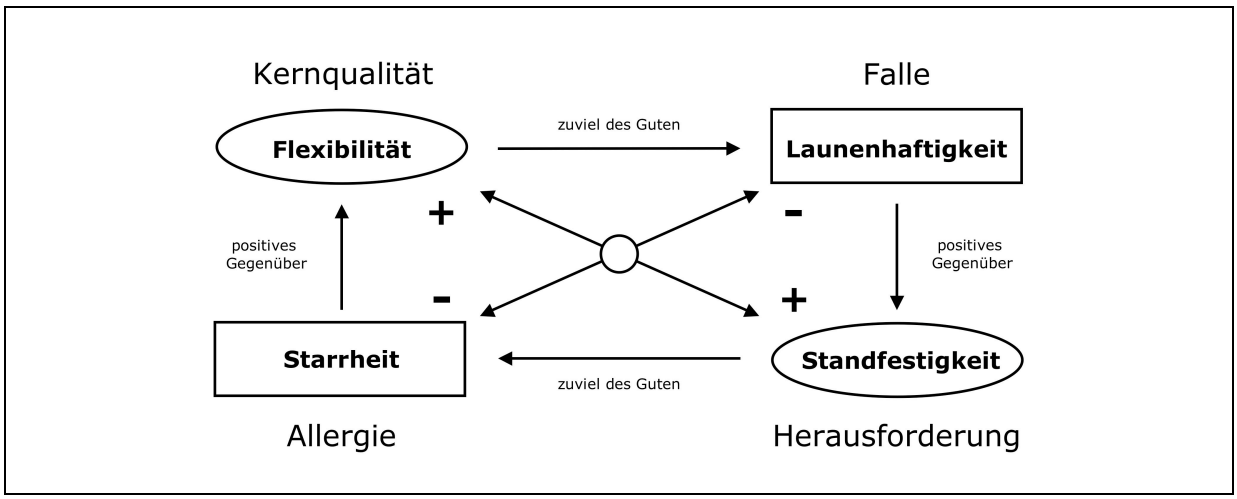


Abbildung 4.3 Darstellung des Kernquadrats (nach Ofman, 2005, S. 40).

Unklar ist, wie es zu dieser doch frappierenden Parallelität zwischen dem Kernquadrat und dem Wertequadrat kam, da Ofman angibt, 1987 ohne Kenntnis des Helwigschen Wertequadrates mit der Entwicklung seines Konzeptes begonnen zu haben. Er schreibt dazu im Vorwort der deutschen Ausgabe: „Wie überrascht war ich, als sich während der Übersetzung dieses Buches herausstellte, dass in Deutschland bereits ein Buch veröffentlich ist [dasjenige von Schulz von Thun], in dem eine Art des Kernquadrates beschrieben wird" (Ofman, 2005, S. 12). Die von ihm verwendeten Ausdrücke „Balance" (S. 39) „zuviel des Guten" oder „entarten" (S. 58) lassen den Einfluss von Helwig jedoch kaum leugnen.

Bevor ich vertiefter auf die praktischen Einsatzmöglichkeiten des Wertequadrates in Diagnostik und Therapie eingehe, stelle ich im nächsten Kapitel dessen historische Wurzeln und die beiden das Wertequadrat definierenden Grundprinzipien Polarität und Entartung dar.

4.2 Geschichte und Grundlagen des Wertequadrates

Die Grundüberlegungen zum Wertequadrat beschrieb Helwig erstmals in seiner 1936 veröffentlichten „Charakterologie" unter den Typologien im Kapitel „Wertideal-Typen". Er veröffentlichte dann 1948 in einem Artikel in der Zeitschrift Psyche eine Fortführung und Präzisierung seiner Gedanken zur Typologisierung des menschlichen Charakters und bezeichnete sein Modell als „Wertequadrat". In der Neuauflage der „Charakterologie" 1951 übernahm er diese Beschreibung und stellte sie an den Beginn des Kapitels zu den Typologien. In Abbildung 4.4 ist die Evolution des Modells von der „Struktur aller Wertebegriffe" (1936) zum Wertequadrat (1948) dargestellt.

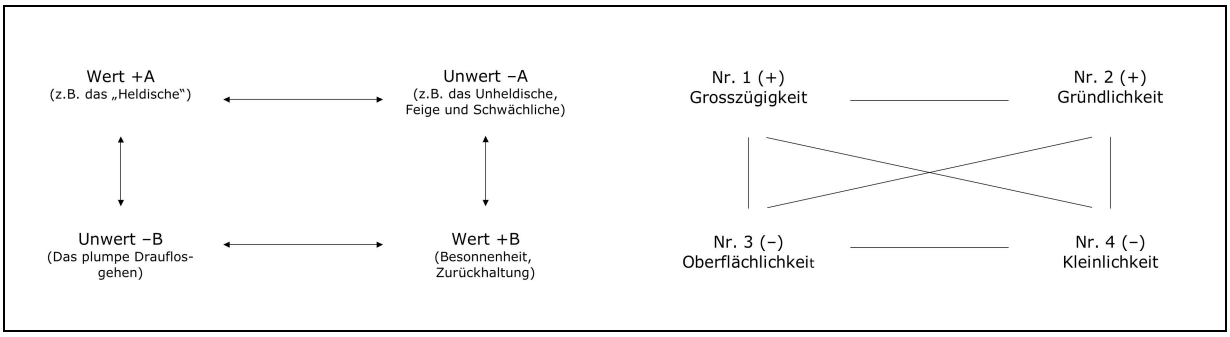


Abbildung 4.4 „Struktur aller Wertebegriffe" (links) und ein Beispiel eines Wertequadrates (rechts) (Helwig, 1936, S. 61 resp. 1948, S. 122).

Mit seinen Überlegungen zum Wertequadrat stand Helwig sowohl inhaltlich wie auch methodisch noch ganz in der Linie der philosophisch orientierten Psychologen, ohne sein Gedankenmodell auf konkrete eigene Beobachtungen oder gar Untersuchungen abzustützen. So führte er in seinem Buch „Charakterologie" (1936) im Kapitel über systematische Typologien auch vorwiegend Schriftsteller und Philosophen auf (Schiller, Nietzsche, Jaspers, Spranger und Jung). Im Anschluss an die Kritik von Sprangers Typologie der Lebensformen formulierte Helwig seine „Wertideal-Typen: Die 'Vierheit' aller Wertbegriffe". Sprangers Typen sind ihm zu rein, das heisst zu einseitig formuliert und stellen in seinen Augen eine „Entartung durch Verabsolutierung" (Helwig, 1936, S. 55) dar. Da seiner Meinung nach „reine" Typen als minderwertige Menschen anzusehen sind, da sie einzelne Werte isoliert verkörpern, stellt für ihn die Idealform eine Balance zwischen zwei Werten dar.

Beim Konzept der Übertreibung oder der „Entartung" lehnte sich Helwig an Jaspers (1919) an, welcher vier Entartungsformen beschreibt, die jeder Menschentyp annehmen kann: Undifferenziertheit, Verabsolutierung, Formalisierung und Unechtheit. Die Idee der Entartung oder Übertreibung lässt sich bis in die Antike zurückverfolgen: In seiner Nikomachischen Ethik unterteilt Aristoteles das Kontinuum der Affekte in ein Übermass, ein Mittleres und einen Mangel und fügt als Beispiel den Umgang mit Geld an: „In Geldsachen, im Geben wie im Nehmen, ist die Mitte Freigebigkeit, das Übermass und der Mangel Verschwendung und Geiz ..." (Aristoteles, 1972, S. 35). Diese drei Eigenschaften sind alle einander entgegengesetzt, wobei die Extreme sowohl im Gegensatz zur Mitte als auch zueinander stehen. Die Mitte, das rechte Mass, teilt Helwig in zwei positive Werte auf, die zueinander in einem ausgehaltenen Spannungsverhältnis stehen.

Wie er auf die Idee mit dem Wertequadrat gekommen ist, beschreibt Helwig nicht. In seinem Buch „Handschrift und Charakter" von 1917 führt Ludwig Klages mit seinen graphologischen Merkmalsbedeutungstabellen jedoch eine Vorform des Wertequadrates ein, wie Abbildung 4.5 zeigt:

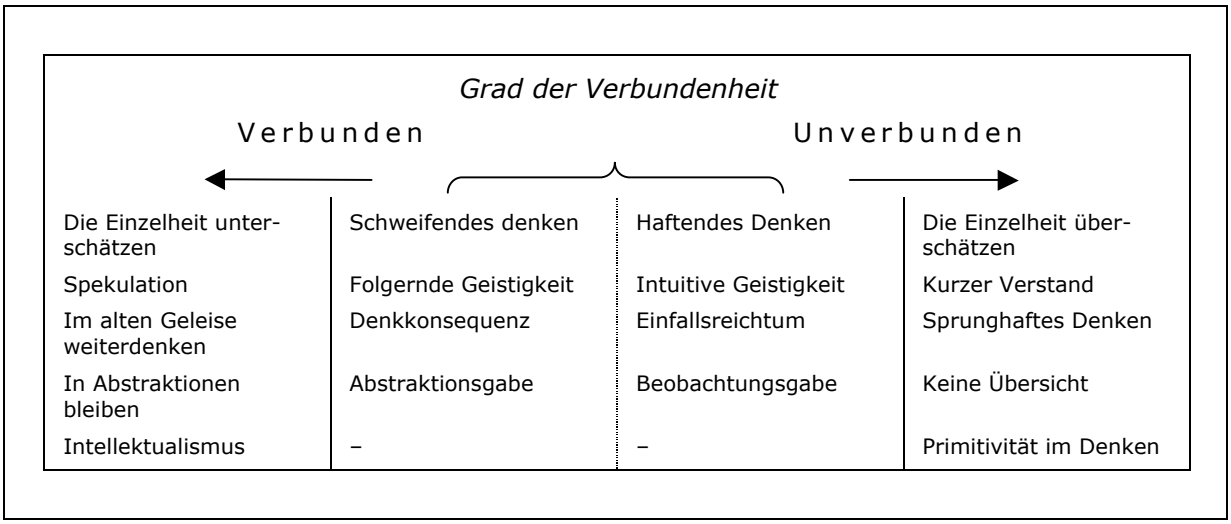
Affektivitätsgrad			
Unaffizierbarkeit		Affizierbarkeit	
+	–	+	–
Gleichmut	Unempfänglichkeit	Empfänglichkeit	Irritabilität
Gelassenheit	Indifferenz	Sensibilität	Reizbarkeit
Beschaulichkeit	Apathie	Zartheit	Empfindlichkeit
Stetheit	Stumpfheit	Eindrucksvermögen	Aufgeregtheit
Ruhe		(Weichheit)	Unruhe
			Launenhaftigkeit
			Streitsucht

Abbildung 4.5 Deutungen des Ebenmasses des Schriftbildes (nach Klages, 1917, S. 16).

Den von Klages erstellten Merkmalsbedeutungstabellen liegt eine Doppeldeutigkeit der graphologischen Merkmale zu Grunde. So teilt er deren Bedeutungen in eine positive und negative Gruppe auf, „je nachdem, ob die Variable als Ausdruck der Stärke einer psychischen Kraft oder der Schwäche einer Gegenkraft zu verstehen ist" (Müller & Enskat, 1993, S. 121). Dabei entscheidet das Formniveau des Schriftbildes, ob das jeweilige Schriftmerkmal positiv oder negativ zu deuten

ist. Schriften mit einem hohen Formniveau wirken auf Grund des starken Rhythmus und dem Eigenartsgrad lebendig und persönlich. Klages sieht in einem hohen Formniveau ein Zeichen von Gestaltungskraft, welche Ausdruck von allgemeiner Lebensintensität und –fülle und somit die Grundlage aller Begabungen sein soll.

Es ist anzunehmen, dass sich Helwig bei der Entwicklung des Wertequadrates stark von Klages inspirieren liess, zumal sich in seiner „Charakterologie" ein ganzes Kapitel dessen Charakterlehre widmet.



Die Einzelheit unterschätzen	Schweifendes denken	Haftendes Denken	Die Einzelheit überschätzen
Spekulation	Folgernde Geistigkeit	Intuitive Geistigkeit	Kurzer Verstand
Im alten Geleise weiterdenken	Denkkonsequenz	Einfallsreichtum	Sprunghaftes Denken
In Abstraktionen bleiben	Abstraktionsgabe	Beobachtungsgabe	Keine Übersicht
Intellektualismus	–	–	Primitivität im Denken

Abbildung 4.6 Deutungen des Grades der Verbundenheit im Schriftbild (nach Wieser, 1960, S. 101).

Eine Zwischenform der Konzepte von Klages und Helwig bilden die für die Betriebsgraphologie entwickelten Bedeutungstabellen von Wieser (1960), indem sie die Merkmalsbedeutungen von Klages in der Struktur eines Wertequadrates darstellt, wobei sie weder diesen Begriff verwendet noch einen Bezug zu Helwig herstellt. Wie Abbildung 4.6 zeigt, unterteilt Wieser in Anlehnung an Klages das Kontinuum der Ausprägung eines Schriftmerkmals in vier Abstufungen, ordnet diese jedoch in zwei gemässigte Mittelkategorien (z. B. Abstraktionsgabe vs. Beobachtungsgabe) und je eine Extremkategorie (z. B. in Abstraktionen bleiben vs. keine Übersicht) ein. Dabei geht sie von Überlegungen aus, die vergleichbar sind mit den von Helwig postulierten, dem Wertequadrat zugrunde liegenden psychischen Mechanismen: „[Es ist] das selbstlos denkende und wollende Ich ..., das ein Abgleiten in jene selbstischen Extreme [– Egoismen –] nicht zulassen wird – Extreme, die psychologisch an den entgegengesetzten Endpunkten ein und derselben Skala lägen" (Wieser, 1960, S. 52). Weiter geht auch sie davon

aus, dass die Merkmale der beiden Mittelkategorien in einem ausgewogenen Verhältnis zueinander auftreten müssen, damit ein Abgleiten in eine der Extremkategorien verhindert wird: „Die Regelmässigkeit [muss] stets auch die Tendenz zu ihrer Auflockerung in sich tragen ..., falls sie noch positiv deutbar sein soll" (Wieser, 1960, S. 79–80).

Das Modell von Helwig fand in der Fachwelt lange Zeit nur wenig Beachtung, bis es 1989 von Schulz von Thun in seinem Buch „Miteinander Reden 2" unter der Bezeichnung „Werte- und Entwicklungsquadrat" einen hohen Bekanntheitsgrad erreichte. Schulz von Thun fügte den Aspekt der Entwicklung hinzu, da sich das Modell besonders gut eignet, um damit auch Vorgänge der zwischenmenschlichen Kommunikation und Persönlichkeitsbildung zu beschreiben, und initiierte so den Einsatz des Wertequadrates in der Beratung (z. B. Kronshage, 1995). Gloor übertrug 1993 inspiriert durch Schulz von Thun das Wertequadrat auf den Bereich der Personalbeurteilung und –selektion und 1995 erschien zum selben Themengebiet die Publikation von Eberle und Hartwich. Danach dauerte es mehr als weitere zehn Jahre bis Westermann 2007 als Herausgeber eine Monografie zum Entwicklungsquadrat veröffentlichte und somit diesem Denkansatz einen festen Platz im Personalmanagement zuwies.

4.3 Vorgehen bei der Entwicklung eines Wertequadrates

Helwig hat in seinen Schriften nie konkret ausgeführt, wie man bei der Entwicklung eines Wertequadrates vorgehen soll. Erst Schulz von Thun (1989) beschreibt dies explizit im Kapitel mit der Überschrift „Wie konstruiere ich ein Wertequadrat". Er schlägt dabei – wie in Abbildung 4.7 dargestellt – zwei mögliche Varianten vor: Bei der ersten beginnt er mit dem positiven Wert und sucht nach dessen entwertender Übertreibung. Danach sucht er nach dem Gegenteil dieser Übertreibung, welches gleichzeitig den positiven Gegenwert zum Ausgangspunkt darstellt. Abschliessend definiert er die Übertreibung dieser zweiten Tugend. Bei der zweiten Variante startet er den Konstruktionsprozess bei einer Untugend und beschreibt anschliessend die (wiederum negative) Übertreibung davon. Er sucht dann einen der beiden positiven Werte, welcher die gute Eigenschaft des einen Unwertes und den Gegensatz der anderen Untugend repräsentiert. Den Schluss bildet die Definition der ausgleichenden Ergänzung des positiven Wertes.

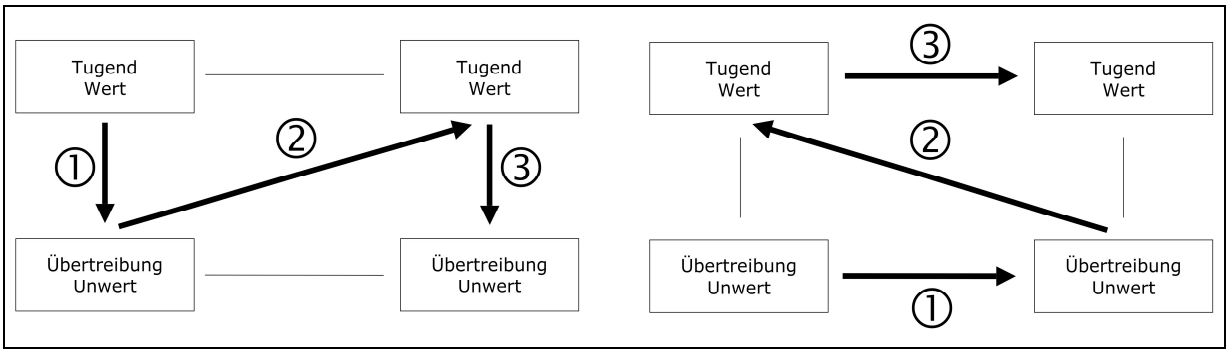


Abbildung 4.7 Vorgehen beim Definieren der vier Wertequadranten nach Schulz von Thun (1989).

Eine andere Vorgehensweise beschreibt Gloor (2007c): Bei der Definition der Wertequadranten nimmt er die eine Tugend als Referenzgrösse. Diese definiert er quasi in Abgrenzung zu deren Gegenteil, also der diagonal gegenüberliegenden Übertreibung. Danach sucht er nach der Übertreibung der gewählten Tugend, um das Wertequadrat schliesslich mit dem Gegengewicht – der Schwesterntugend – zu komplettieren.

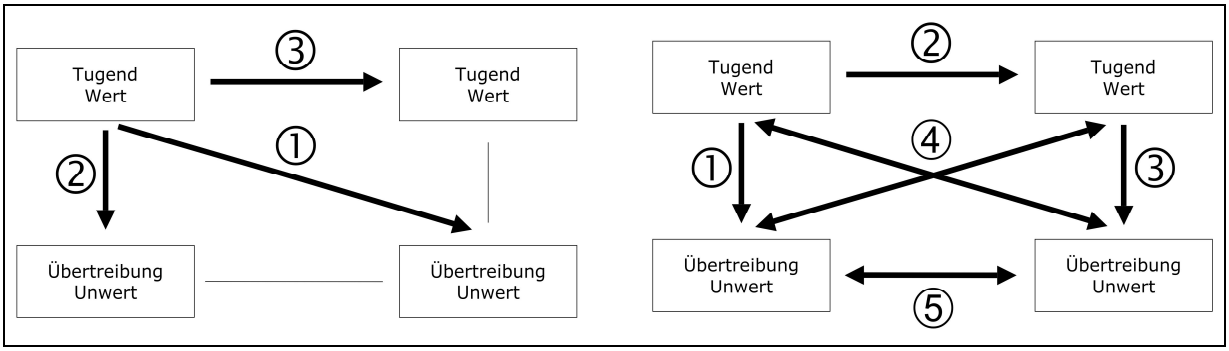


Abbildung 4.8 Vorgehen beim Definieren der vier Wertequadranten nach Gloor (2007c) und Westermann (2007b).

In Abbildung 4.8 habe ich neben dem Vorgehen von Gloor auch noch die von Westermann (2007b) beschriebene Methode dargestellt. Dieser geht dabei von den Kernaussagen zum Wertequadrat – von ihm als Denkschritte des Entwicklungsquadrates bezeichnet – aus: Mit dem Grundgedanken des Wertequadrates „Verhalten ist immer relativ. Jede Stärke kann sich in eine Schwäche verwandeln, wenn des Guten zuviel getan wird." (S. 11) definiert er ausgehend von einer Tugend deren Übertreibung. Anschliessend gelangt er über die Gegentugend zu deren Übertreibung: „Die polare Gegentugend ist erforderlich als kom-

plementäres zweites Spielbein zur Vermeidung von kritischen Engpässen bzw. Gleichgewichtsstörungen. ... Auch diese polare Gegentugend birgt in sich die Gefahr des Überzeichnens" (S. 11 und 12). Schliesslich folgen als vierte und fünfte Denkschritte die vollständige Verknüpfung der vier Wertequadranten: „Als Entwicklungshinweis dient das schräg gegenüberliegende Stärkefeld." (S. 13) und „Überkompensation als Pendeln zwischen zwei Extrempolen ohne echte Lebenslösung." (S. 14).

Da die Beziehungen der vier Wertequadranten untereinander vollständig definiert sind und somit das Wertequadrat als Gesamtes in sich stimmig sein muss, sind wohl alle vier vorgestellten Vorgehensweisen zur Entwicklung eines Wertequadrates als idealtypische Abläufe zu verstehen. Im konkreten Fall wird es so sein, dass der Konstrukteur die Einzelschritte, aber auch den Gesamtprozess mehrmals und in unterschiedlichen Reihenfolgen durchläuft, bis das Endprodukt steht. Zudem hängt es auch vom gewählten Tugendbegriff ab, welcher der drei anderen Wertequadrant-Begriffe den Konstrukteur auf Grund seiner Offensichtlichkeit geradezu anspringt und somit den Ablauf der weiteren Konstruktionsschritte vorgibt. Im folgenden Kapitel gehe ich darauf ein, wie auf der Grundlage der beschriebenen Überlegungen erstellte Wertequadrate als Dimensionsbeschreibungen in der Persönlichkeitsdiagnostik Eingang gefunden haben.

4.4 Das Wertequadrat als Methode in der Persönlichkeitsdiagnostik

Helwig (1948) sah im Wertequadrat einen „gedanklichen Kunstgriff" (S. 121) zur Darstellung der inneren Zusammenhänge von psychologischen Werten und zur präzisen Beschreibung von Persönlichkeitseigenschaften. Er blieb zeitlebens seiner philosophisch-erklärenden Denkweise treu und lieferte in seinen Schriften keine Ideen oder Anregungen, wie praktisch tätige Psychologen sein Modell im Alltag nutzbar machen könnten. D'Heureuse (1951) erwähnte zwar in ihren „Gedanken zum Wertequadrat", dass dieses „die Stellung einer Diagnose und die Therapie ... seelischer Gleichgewichtsstörungen" (S. 119) erleichtere, führte dies jedoch nicht weiter aus. Erst Schulz von Thun (1989) zeigte anhand einer Vielzahl anschaulicher Beispiele auf, wie das Wertequadrat bei der Analyse und Überwindung von zwischenmenschlichen Kommunikationsproblemen in der Praxis eingesetzt werden kann. Dabei liegt der grosse Vorteil dieses Konzeptes darin, dass es einen Erklärungsansatz für scheinbar widersprüchliches Handeln bietet (Birkhan, 1998).

Diese Qualität des Wertequadrats prädestiniert es denn auch für den Einsatz in Unternehmen, da – wie Neuberger (1983, 1995, 2002) aufzeigt – das dem Wertequadrat zugrunde liegende Prinzip der Polarität auch in der Personalführung seine Gültigkeit hat: Ausgehend von den Michigan- und Ohio-Studien zum Führungsverhalten (Aufgaben- und Mitarbeiterorientierung, Kahn & Katz, 1953 resp. Fleishman, 1953; Halpin & Winer, 1957; Hemphill & Coons, 1957) listet er unter dem Titel „Führen als widersprüchliches Handeln" 14 Rollendilemmata der Führung auf, wie zum Beispiel Kontrolle und Vertrauen oder Herausforderung und Fürsorge (siehe auch Abbildung 4.9). Dabei weist er darauf hin, dass die jeweiligen Polaritäten eines Dilemmas nicht unbedingt als die Endpunkte eines Kontinuums anzusehen sind, sondern dass es sich auch um voneinander unabhängige Dimensionen handeln kann. „Die innere Zwiespältigkeit des Führens fordert Kompromisse zwischen Alternativen, die jeweils *beide* unverzichtbar sind. Die völlige Vernachlässigung eines Aspekts würde mit Sicherheit das Scheitern als Vorgesetzter bedeuten" (Neuberger, 1983, S. 24). Die Nähe dieser Überlegungen zu denjenigen von Helwig ist offensichtlich. Dies hat Blickle (1993) dazu angeregt, die Führungsdilemmata unter Verwendung des Konzeptes der polaren Integration von Groeben (1981) in Wertequadrate einzubetten, womit er die produktiven und konstruktiven Kräfte dieser scheinbar ausweglosen Führungssituationen aufzeigen konnte.

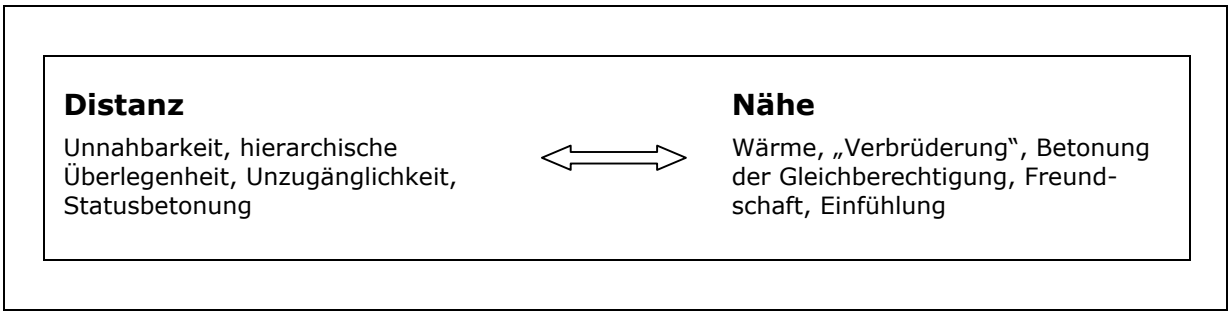


Abbildung 4.9 Beispiel eines Rollendilemmas der Führung (Neuberger, 1995, S. 91).

Gloor hat das enorme Potenzial des Wertequadrats für den Einsatz in der Persönlichkeitsdiagnostik im Rahmen der Personalselektion erkannt und – zumindest im Bereich der Assessment Center – populär gemacht. In seinem 1993 erschienen Buch „Die AC-Methode" bildet das Wertequadrat das „Star-Werkzeug" (S. 13) für die Umsetzung von Anforderungsprofilen in Beobachtungs- und Beurteilungsbogen. Den Hauptgewinn sieht Gloor dabei in der Überwindung der realitätsfremden Polarität zwischen gut und nicht gut, indem er stattdessen von „Schwesterntu-

genden" (S. 14) spricht, welche in einer dialektischen Ergänzung zueinander stehen.

Abbildung 4.10 zeigt einen Ausschnitt aus dem Beobachtungsbogen von Gloor. Dieser ersetzte die Wertbegriffe, wie sie Helwig und Schulz von Thun verwendeten, durch Verhaltenskategorien und operationalisierte diese, indem er sie mit typischen Handlungsbeispielen ergänzte. Gloor (2007b) setzte das Wertequadrat zusätzlich für die Formulierung von firmenspezifischen Führungsgrundsätzen ein (von ihm als „Radarschirme" bezeichnet), wobei er die üblicherweise abstrakt, althergebracht und moralisch formulierten Grundsätze mit Hilfe des Wertequadrates konkretisierte und somit gleichzeitig zum wünschbaren auch das defizitäre Führungsverhalten aufzeigen konnte. Zur Arbeit mit dem Beurteilungsbogen gibt Gloor (1993) folgende Anweisung:

> Diese Skala darf nicht als Notenskala verwendet werden; die Zahlen dieser Skala repräsentieren nicht numerische Werte von 1 bis 8, sondern Standorte auf der Bandbreite des Itembalkens. Es wäre also absolut nicht im Sinn und Geist dieses Instrumentes, wenn Sie Ihre Bewertungszahlen am Schluss zusammenzählen und deren Durchschnitt berechnen würden. (S. 150)

Gloor empfiehlt damit einen Mittelweg zwischen einer unstrukturierten, eher klinisch orientierten und einer vollstrukturierten, mathematisch-statistischen Auswertung, wobei er nicht ausführt, wieso er die Bildung von Summenwerten innerhalb einer – in einem Assessment Center mehrfach erhobenen – Persönlichkeitsdimension als unzulässig erachtet.

	❶ ❷	❸ ❹	❺ ❻	❼ ❽
Kontakt- und Begegnungs-stil	**überfahrend:** monologische Redseligkeit / wirkt vorlaut, selbstgefällig, geschwätzig / bringt am liebsten sich selbst zur Aufführung	**Kontaktbereitschaft:** stellt von sich aus spontan Kontakt her / wirkt mitteilungsfreudig, farbig, belebend, lebendig, initiativ	**Zurückhaltung:** abwartend / vorsichtig / eher reaktiv als aktiv	**sich entziehend:** wortkarg / fad / zaghaft / langweilig / zugeknöpft / einsilbig

Abbildung 4.10 Ausschnitt aus dem Beobachtungsbogen von Gloor (1993, S. 145).

Später unternahm Gloor (2007a) den Versuch, das Wertequadrat und die Big Five der Persönlichkeit in seinem so bezeichneten „BFWQ-Modell" zu verbinden, da er sich daran störte, dass Big-Five-Skalen immer aus zwei Gegenpolen bestehen, welche sich gegenseitig ausschliessen. Für ihn ist dies ein Beispiel „der traditionsreichen Erstarrtheit der Entweder-oder-Polarisierung" welche im Gegensatz zur „Dialektik der dynamischen Sowohl-als-auch-Balance" steht: „Warum will das Big-Five-Konzept nicht wahrhaben, dass es Menschen gibt, die *sowohl* offen sind für neue Erfahrungen *als auch* skeptisch oder gar ablehnend gegenüber Neuem – dann etwa, wenn es sich bei diesem Neuen um irgendwelche alberne Trends und Moden handelt?" (Gloor, 2007a, S. 36). Zudem bemängelt er, dass der positive Pol ausführlicher beschrieben wird, als der negative, was er mit dem fehlenden Mut, auch die bei jedem Menschen vorhandenen negativen Persönlichkeitseigenschaften direkt anzusprechen, in Verbindung bringt. Bei der Übertragung der Big Five ins Wertequadrat definierte Gloor zuerst die Schwesterntugenden und Übertreibungen, operationalisierte diese anschliessend mittels beobachtbarer Verhaltensweisen und stellte diese in einem Beobachtungs- und Beurteilungsbogen zusammen. Bei der Beurteilung des beobachteten Verhaltens schlug Gloor – getreu der Grundidee des Wertequadrates, dass jeder Mensch immer von beiden Tugenden etwas in sich trägt – die Setzung von zwei Wertungen pro Dimension vor; es wird also zum Beispiel sowohl die Ausprägung in der Tugend „Extraversion" wie auch in der Tugend „Introversion" festgehalten. Abbildung 4.11 zeigt die Operationalisierung der Dimension „Extraversion" im Beurteilungsbogen von Gloor.

Extraversion			
-10 -5	0 +5	+10 +5	0 -5 -10
Der Hobby-Shower: wirkt bemüht locker und lässig; redet zu allem mit, aber oberflächlich; Attitüde „Ich bin ein toller Hecht"; wirkt laut, grell, grossmaulig, aufdringlich	**extravertiert:** heiteres, lebensfreudiges Naturell; stellt leicht Beziehungen her; ist unterhaltend, steht gerne im Mittelpunkt; wirkt unbekümmert und selbstsicher	**introvertiert:** wirkt zurückhaltend, ruhig, aufmerksam; wirkt in seinen Mitteilungen sachlich; bleibt im Hintergrund / drängt sich nicht vor; kann innehalten und nachdenken	**Die graue Maus:** zögerndes, sich verdeckendes, zurückgezogenes Naturell; wirkt fad, langweilig, verschlossen; kann in einer Gruppe übersehen werden; äussert sich kaum

Abbildung 4.11 Weiterentwicklung des Beobachtungs- und Beurteilungsbogens (Gloor, 2007a, S. 40).

Eberle und Hartwich (1995) nutzten das Wertequadrat als Rückgrat ihrer nach dem Polaritätsprinzip gebildeten „Komplementären Einschätzhilfen" (Abbildung 4.12), welche die Grundlage für ein System zur „Schnelleinschätzung" der Persönlichkeit (Eberle, 2007, S. 50) und für ein Interviewverfahren bilden. Die Idee dazu stammt aus der positiven Psychotherapie und dem differenzierungsanalytischen Inventar von Peseschkian (1977), welcher als Ausgangspunkt für therapeutische Interventionen die für die betroffene Person positiven Aspekte des gesellschaftlich als negativ oder unangemessen erachteten Verhaltens beleuchtet. In diesem Sinne müssten Persönlichkeitsinventare gemäss Eberle und Hartwich so aufgebaut sein, dass die beurteilte Person mit der vorgenommenen Einstufung einverstanden ist und sich auch darüber verständigen kann. Dazu ist eine allgemeinverständliche Sprache zu verwenden und ein Einschätzungssystem einzusetzen, welches Unterschiede zwischen den Menschen und deren Verhalten nachvollziehbar abbilden kann. Zudem muss der Diagnostiker „für das Persönlichkeitsinventar eine wertschätzende Ausdrucksform wählen, welche die charakterologische Einordnung für die Menschen erträglich macht und nicht bereits durch die Art der Formulierung zu Verhaltensblockaden und Erkenntnissperren führt" (Eberle & Hartwich, 1995, S. 88). Damit vertreten sie eine grundsätzlich andere Haltung als Gloor (2007a), welcher diese allzu wohlwollenden Beschreibungen als „Ducksen, Mogeln, Schönreden und Verschleiern" (S. 35) bezeichnet.

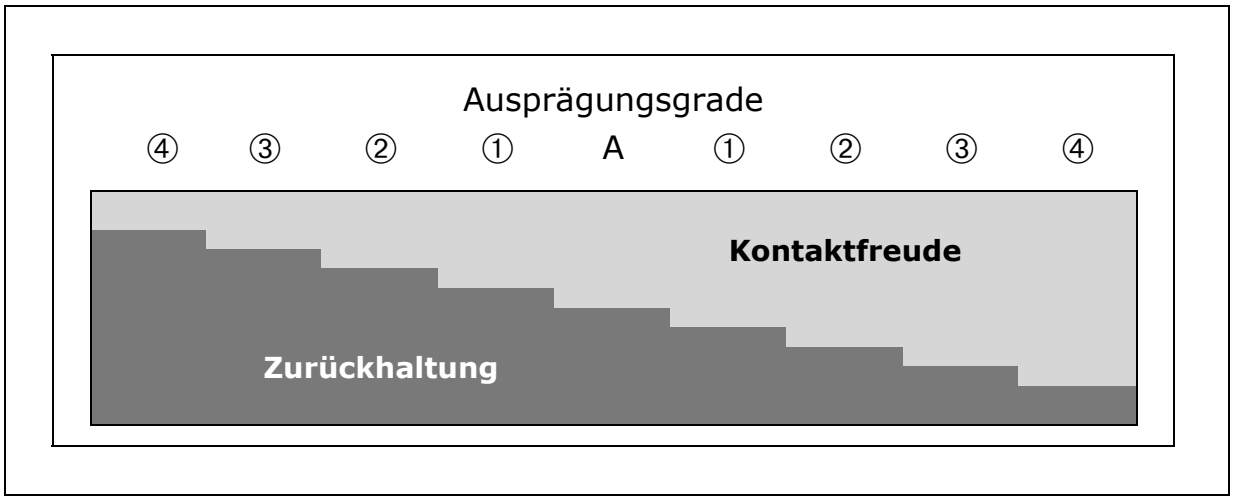


Abbildung 4.12 Ausprägungsgrade im Verhaltenskontinuum (nach Eberle & Hartwich, 1995, S. 95).

Eberle und Hartwich setzen bei ihrem Persönlichkeitsinventar das Polaritätsprinzip ein, davon ausgehend, dass bei jedem Menschen immer beide Elemente einer Polarität komplementär verhaltenswirksam sind. Gemäss der Auffassung,

dass sich menschliches Verhalten im Spannungsfeld jeweils zweier entgegengesetzter Eigenschaften abspielt, kann jedes Verhalten grundsätzlich aus zwei Blickwinkeln betrachtet werden. „Die Konsequenz [daraus] ... ist, dass man in jedem Menschen Stärken entdecken kann und dass jede Eigenschaft eines Menschen positiv formuliert werden kann" (Eberle & Hartwich, 1995, S. 94). Dies zeigt sich schon an ihrer Definition der Pole, welche immer positiv formuliert sind. Als Beispiel seien diejenigen des Gesellungsverhaltens aufgeführt:

> Der zurückhaltende Mensch liebt die Einsamkeit und ist gerne ungestört und für sich allein. Er will sich niemandem aufdrängen, und die Intimsphäre eines Menschen ist ihm heilig. Er liebt die Beobachterrolle am Rande des Geschehens.
>
> Der kontaktfreudige Mensch sucht die Nähe der Menschen. Er will Freud und Leid mit anderen teilen und fühlt sich im grossen Kreise wohl. Er ist sehr umgänglich, geht von sich aus auf andere Menschen zu und ist auch schnell mit ihnen per du. (Eberle & Hartwich, 1995, S. 116–117)

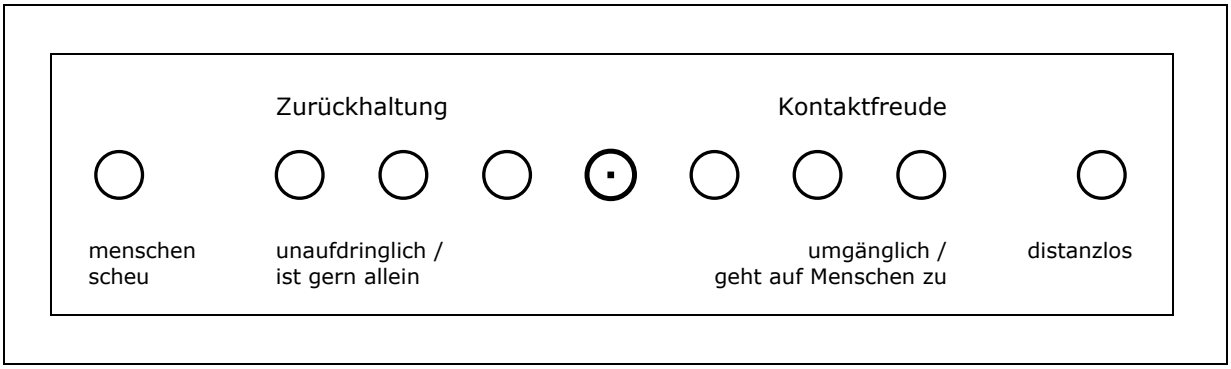


Abbildung 4.13 Darstellung des Verhaltenskontinuums im Einschätzungsbogen des KEH-Systems (nach Eberle & Hartwich, 1995, S. 110).

Zur Einstufung der verschiedenen Persönlichkeitseigenschaften haben Eberle und Hartwich das Verhaltenskontinuum in neun Ausprägungsgrade unterteilt – je vier Stufen auf jeder Polseite und eine Mittelposition, bei welcher sich die Verhaltenstendenzen in einem Gleichgewicht befinden. Personen, welche sich dem mittleren Ausprägungsgrad A zuordnen lassen, zeichnen sich somit dadurch aus, dass sie keine der gegenseitigen Verhaltensweisen bevorzugen und sich situativ richtig verhalten. Auch Personen mit dem Ausprägungsgrad 1 beherrschen beide Verhaltenspolaritäten, ziehen jedoch eine Seite vor. Beim Ausprägungsgrad 2 liegt schon eine starke Tendenz vor, eine Seite deutlich zu bevorzugen. Die Unterentwicklung der gegenseitigen Verhaltenstendenz führt dabei in gewissen Situati-

onen zu gelegentlichem Fehlverhalten. Diese Tendenz verstärkt sich beim Ausprägungsgrad 3, was zu einer grossen Einseitigkeit und zu einer Geringschätzung der unterentwickelten Zwillingseigenschaft führt. Beim Ausprägungsgrad 4 ist die Grenze zur psychischen Krankheit erreicht, wobei „die Einseitigkeit ... als ganz ausgeprägte, manchmal schon geniale Stärke zu erkennen" ist (Eberle & Hartwich, 1995, S. 99). Abbildung 4.13 zeigt, wie die Autoren diese Abstufung in den komplementären Einschätzungshilfen umsetzen.

Briefs (2007) setzt für die Selektion von Führungskräften eine siebenstufige Skala ein (siehe Abbildung 4.14), bei welcher sich jeweils zwei positive Verhaltensweisen gegenüber stehen und jeder Pol zusätzlich durch eine negative Überhöhung markiert wird. Diese komplementär, sich ergänzende Darstellung der Anforderungsdimensionen dient als Grundlage für die Ausformulierung von Interviewfragen und Entwicklung von Kurzübungen. Als besonderen Vorteil dieses Vorgehens sieht er die wertneutrale Erfassung eines breiten Spektrums an Verhaltensweisen sowie die Möglichkeit, positive Aussagen über das Verhalten der eingeschätzten Person machen zu können. Konkret setzt er dies anhand von Interviewfragen um, die beide Verhaltensweisen ansprechen, wobei der Bewerber sein Verhalten einer davon zuordnet. Zur Dimension „in Details denkend – in Zusammenhängen denkend" könnte eine Interviewfrage dann lauten: „Worauf kommt es bei der Delegation an, auf das genaue, detaillierte Beschreiben, was getan werden muss, oder auf die Vermittlung des Zusammenhangs, in dem die Aufgabe steht?" (Briefs, 2007, S. 77).

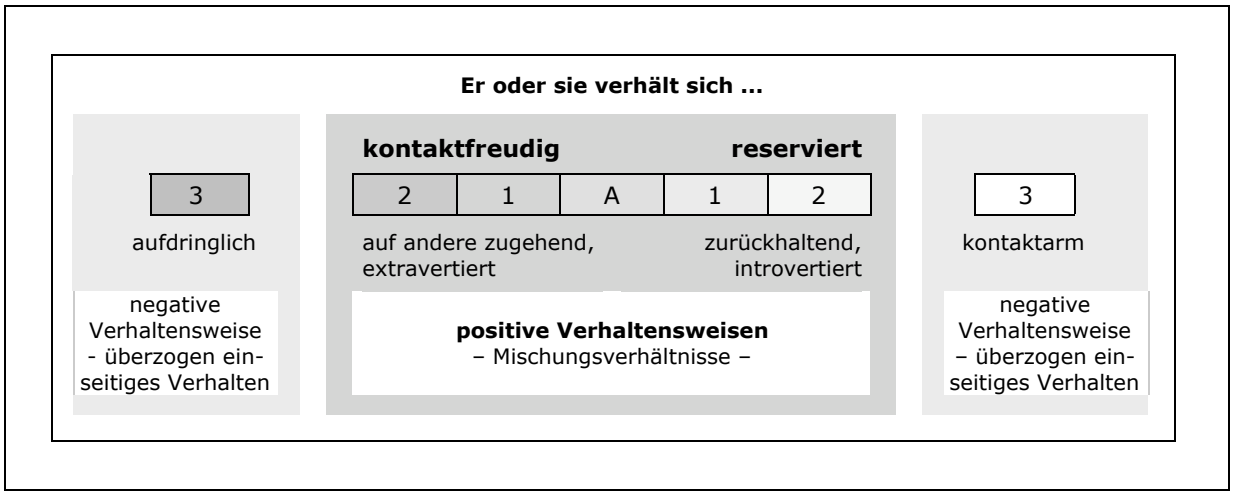


Abbildung 4.14 Skala zur Festlegung von Ausprägungen (nach Briefs, 2007, S. 73).

Schon bei der Personalselektion ist es wichtig, eine möglichst wertfreie Beurteilung vornehmen zu können – bei der Personalentwicklung ist dies ein absolutes Muss. Aus diesem Grund eignet sich das Werte- oder eben Entwicklungsquadrat besonders gut für diesen Einsatzzweck, da es eine positive und primär an den Stärken des Kandidaten orientierte Rückmeldung erlaubt (siehe Birkhan & Reitzig, 2007) und da es die widersprüchlichen Anforderungen an Manager aufzeigt (siehe Neuberger, 1983). Bei der Formulierung von Interviewfragen lassen sich zudem Aspekte der Situationsabhängigkeit gut einbinden. Birkhan und Reitzig (2007) beschreiben für ihre „persönliche Standortbestimmung" jede Anforderungsdimension mit zwei positiv formulierten Eigenschaften (Tugenden). So wählen sie zum Beispiel als Pole des Gestaltungswillens das „Streben nach Freiräumen" und das „Anpassen an Strukturen". Weiter formulieren sie für die Tugenden und Übertreibungen im Wertequadrat entsprechende Definitionen:

Definition: Gestaltungswille ist einerseits das Streben nach Freiräumen und deren Ausfüllen durch eigenes Handeln und andererseits das zielführende Agieren innerhalb vorgegebener Strukturen.

Tugend 1: agiert zielführend innerhalb vorgegebener bzw. selbstgeschaffener Strukturen.

Übertreibung 1: benötigt präzise Vorgaben, Strukturen und Hinweise für das eigene Handeln.

Tugend 2: sucht sich bzw. schafft sich selbstständig Freiräume für das eigene Handeln.

Übertreibung 2: unangepasst, handelt eigenwillig, lehnt Zielvorgaben ab; definiert seine eigenen Spielregeln.

Situationsadäquate Angemessenheit: kann sich an sinnvolle, vorgegebene Strukturen anpassen und schafft sich innerhalb dieser den optimalen Handlungsfreiraum. (Birkhan & Reitzig, 2007, S. 87–88).

In der neunstufigen bipolaren Skala (siehe Abbildung 4.15) steht die Skalenmitte (5) für die Fähigkeit, „beide Eigenschaftsvarianten in situationsangemessener Weise anwenden zu können. Die Extrempole 1 und 9 stellen die jeweiligen dysfunktionalen Übertreibungen dar" (Birkhan & Reitzig, 2007, S. 86).

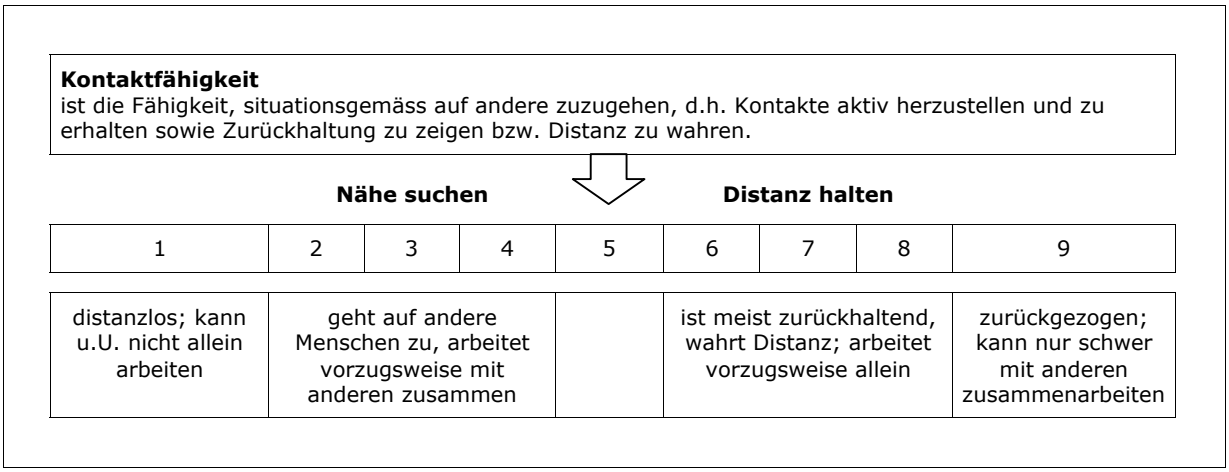


Abbildung 4.15 Bipolare Skala nach Birkhan (1998; S. 166, resp. 2007, S. 26).

4.5 Das Wertequadrat als Konstruktionsprinzip zur Reduktion des Gebens verfälschter Antworten in Persönlichkeits-Fragebogen

Auch wenn bis heute noch keine empirischen Befunde für die von Helwig (1948) formulierten psychodynamischen Erklärungsansätze vorliegen – und diese wahrscheinlich auch nie erbracht werden können – so belegen die oben aufgeführten Anwendungsbeispiele des Wertequadrates die praktische Nützlichkeit dieses Denkansatzes bei psychologischen Interventionen und in der Persönlichkeitsdiagnostik. Dabei ist es vor allem die Möglichkeit der wertfreien Beschreibung der Persönlichkeit, welche das Wertequadrat für den Einsatz in der Personalselektion prädestiniert. Damit lässt sich die heute noch weit verbreitete Vorstellung überwinden, dass ein Kandidat umso besser geeignet ist, je stärker ausgeprägt bei ihm eine für eine bestimmte Position geforderte Persönlichkeitseigenschaft ist. Auf dieser Annahme basieren implizit noch praktisch alle Persönlichkeitsinventare, welche sich aus unipolaren Skalen zusammensetzen. Dabei ist der entscheidende Negativaspekt beim Einsatz von Persönlichkeits-Fragebogen in der Personalselektion ihre leichte Durchschaubarkeit, welche durch den bei unipolaren Skalen häufig zugrunde liegenden Je-mehr-desto-besser-Ansatz begründet ist. Somit ist es für die Bewerber ein Leichtes, den Test „zu durchschauen" und sich bei dessen Bearbeitung beschönigt darzustellen (*Impression Management*) oder gar unzutreffende Angaben abzugeben (*Faking*) (Rothstein & Goffin, 2006; Tett & Christiansen, 2007. Einen umfassenden Überblick zum Phänomen Faking bieten Griffith und Peterson (2006) in ihrer Monografie.).

Viele Studien belegen denn auch, dass sich Bewerber in Persönlichkeits-Fragebogen beschönigt darstellen (Barrick & Mount, 1996; Birkeland, Manson, Kisamore, Brannick & Smith, 2006; Donovan, Dwight & Hurtz, 2003; Griffith, Chmielowski & Yoshita, 2007; Rosse, Stecher, Miller & Levin, 1998; Viswesvaran & Ones, 1999). Da ein Teil der Studien experimentellen Charakter hat, ist unter Wissenschaftlern jedoch noch umstritten, ob dies in der realen Bewerberselektion auch ein Problem darstellt, in dem Sinne, dass Personalfachleute durch den Einsatz von Persönlichkeits-Fragebogen falsche oder suboptimale Entscheide treffen (Converse, Peterson & Griffith, 2009; Hogan, Barrett & Hogan, 2007; Hough, Eaton, Dunnette, Kamp & McCloy, 1990; Mueller-Hanson, Heggestad & Thornton, 2003; Ones, Viswesvaran & Reiss, 1996; Rosse et al., 1998). Die unklare Forschungslage führte in der Vergangenheit auch immer wieder zu heftig geführten Diskussionen unter Experten, ob der Einsatz von Persönlichkeits-Fragebogen im Rahmen der Personalselektion nutzbringend und zulässig ist (Morgeson, Campion, Dipboye, Hollenbeck, Murphy & Schmitt, 2007a, 2007b; Ones, Dilchert, Viswesvaran & Judge, 2007; Tett & Christiansen, 2007).

Mittels verschiedener Vorgehensweisen versuchten Fachexperten, das Ausmass des Verfälschens von Antworten in Persönlichkeits-Fragebogen zu reduzieren respektive zu kontrollieren, wobei viele Ansätze die in sie gesetzten Erwartungen nicht erfüllen konnten. So zeigte sich, dass sich mit der Einschränkung der Antwortzeit Faking nicht verhindern lässt (Holden, Wood & Tomashewski, 2001) obwohl das verzögerte Geben der Antwort in einem Persönlichkeits-Fragebogen unter bestimmten Rahmenbedingungen ein Hinweis auf Faking ist (Holden & Hibbs, 1995; Vasilopoulos, Reilly & Leaman, 2000). Grosse Erwartungen setzten die Experten in Skalen zur Erfassung des sozial erwünschten Antwortverhaltens, deren Ergebnisse sie zur Korrektur der Werte der Persönlichkeitsskalen heranzogen. In keiner Studie konnte jedoch der erhoffte Effekt nachgewiesen werden (Christiansen, Goffin, Johnston & Rothstein, 1994; Ellingson, Sackett & Hough, 1999; Griffith & Peterson, 2008; Hough, 1998; Li & Bagger, 2006; Ones et al., 1996; Piedmont, McCrae, Riemann & Angleitner, 2000; Schmitt & Oswald, 2006). Vielversprechend sind hingegen neuere Ansätze zur Aufdeckung und Korrektur von verfälschten Angaben in Persönlichkeits-Fragebogen, wie der Einsatz der Item Response Theory (Holden & Book, 2009; Zickar, Gibby & Robie, 2004) oder die statistische Analyse des Antwortverhaltens von für Faking besonders anfälligen Items, wie zum Beispiel die *Employment-Related Motivation Distortion Predictive Scale* (Hakstian & Ng, 2005) oder der *Covariance Index* (Christiansen, 2008). Uneindeutig sind die Effekte von (Warn-)Instruktionen, welche den Testbearbeiter darauf hinweisen, dass Verfälschungen aufgedeckt werden können: Einzelne Forscherteams konnten positive Effekte nachweisen (z. B. Dwight &

Donovan, 2003; Goffin & Woods, 1995; McFarland, 2003) andere nicht, respektive nur unter bestimmten Bedingungen (z. B. Converse, Oswald, Imus, Hedricks, Roy & Butera, 2008; Robson, Jones & Abraham, 2008; Vasilopoulos, Cucina & McElreath, 2005).

Da die bisher dargestellten Möglichkeiten zur Verhinderung oder Korrektur von Antwortverfälschungen keine oder nur geringfügige Effekte zeigen, noch nicht ausgereift oder genügend untersucht sind, bleibt nur noch die Art und Weise der Konstruktion des Testverfahrens, mit welcher dem Problem begegnet werden könnte. Schon in den Fünfzigerjahren des letzten Jahrhunderts setzten Testkonstrukteure deshalb das Forced-Choice-Antwortformat ein (z. B. Edwards, 1954). Hierbei sind die Items aus zwei oder vier Aussagen aus unterschiedlichen Skalen jedoch mit gleich eingestufter sozialer Erwünschtheit zusammengesetzt, wobei der Testbearbeiter zu entscheiden hat, welche der Aussagen auf seine Person am ehesten (oder auch am wenigsten) zutrifft. Da der Entscheidungsprozess bei der Bearbeitung von Forced-Choice-Aufgaben einen Einfluss auf die psychometrischen Eigenschaften des Testverfahrens hat (Meade, 2004) und das Format messtheoretisch auf Grund der sogenannten ipsativen Messung mit verschiedenen Problemen behaftet ist, raten einige Wissenschaftler generell vom Einsatz dieser Methode ab (z. B. Cornwell & Dunlap, 1994; Hicks, 1970; Johnson, Wood & Blinkhorn, 1988; Travers, 1951). Zudem konnten die Testkonstrukteure in den Fünfziger- und Sechzigerjahren auch keine grössere Robustheit dieses Antwortformates gegenüber Faking nachweisen (Graham, 1958; Scott, 1963; aber auch Heggestad, Morrison, Reeve & McCloy, 2006; Lammers & Frankenfeld, 1999), was dazu führte, dass man sich lange Zeit nicht mehr wissenschaftlich damit befasste (Rothstein & Goffin, 2006).

Erst Mitte der Neunzigerjahre setzte die Forschungstätigkeit dazu wieder ein, nachdem SHL, einer der weltweit grössten Testanbieter, seinen Persönlichkeitstest in einer Forced-Choice-Variante anbot und mit eigener Forschung die Überlegenheit seines Verfahrens belegten (Baron, 1996; Bartram, 1996, 2007. Siehe auch Bowen, Martin & Hunt, 2002; Christiansen, Burns & Montgomery, 2005; Hirsh & Peterson, 2008; Jackson, Wroblewski & Ashton, 2000). Offenbar stellt das Forced-Choice-Format nur ein Problem dar, wenn ein Testverfahren wenige Skalen umfasst, weil dann die einzelnen Skalen nicht mehr unabhängig voneinander sind. Bowen et al. (2002) konnten in einer Validierungsstudie nachweisen, dass dieses Problem bei einer genügend grossen Anzahl von Skalen (in ihrem Fall 30) nicht mehr von Bedeutung ist. Es zeigte sich zudem, dass bei der Bildung der Forced-Choice-Tetraden die Ausprägung der sozialen Erwünschtheit eines Items nicht – wie üblich – allgemein eingestuft werden darf, da bei der

Personalselektion die Erwünschtheit oder Angemessenheit eines Verhaltens vom Anforderungsprofil der zu besetzenden Stelle abhängt (Rothstein & Goffin, 2000).

Jackson et al. (2000) setzten bei der Entwicklung ihrer Integritäts-Skala die *dichotomous quartet method* (Dunnette, McCartney, Carlson & Kirchner, 1962) ein, bei welcher zwei sozial erwünschte Aussagen mit zwei sozial unerwünschten kombiniert werden und somit Items erzeugt werden, welche eine entfernte Ähnlichkeit mit Wertequadraten haben. Die Fragebogen-Bearbeiter müssen dabei die Aussage wählen, welche ihr Verhalten am besten und diejenige, welche es am wenigsten gut beschreibt. Damit wollen die Testautoren verhindern, dass sie bei den Bearbeitern Reaktanz erzeugen, wenn diese eine von zwei unerwünschten Verhaltensweisen als für sie zutreffend wählen müssen. Für die Bestimmung der sozialen Erwünschtheit nahmen die Autoren die durchschnittliche Zustimmungsrate und legten die Items 20 Personen vor, welche auf einer siebenstufigen Skala angeben mussten, wie stark dieses Item einen Bewerber motiviert, durch seine Antwort einen guten Eindruck zu hinterlassen. Anhand eines Computerprogramms nahmen sie dann die Zuordnung der Item-Dyaden vor (gleiche Zustimmungsrate, gleiche soziale Erwünschtheit) und bildeten abschliessend die Tetraden, indem sie eine hoch erwünschte mit einer nichterwünschten Dyade zusammengefassten. Die experimentelle empirische Überprüfung zeigte, dass dieses Forced-Choice-Format wesentlich robuster gegenüber Faking ist, als eine herkömmliche, likert-skalierte Persönlichkeitsskala.

Forced-Choice-Verfahren haben gegenüber traditionellen, likert-skalierten Persönlichkeits-Fragebogen jedoch auch ein paar Nachteile: So akzeptieren die Testbearbeiter diese in der Regel schlechter, weil es oftmals mühsam ist, sich zwischen zwei gleichwertigen Verhaltensweisen entscheiden zu müssen. Zudem geben sie an, dass sie sich bei Likert-Skalen besser beschreiben können, diese einfacher zu bearbeiten sind und somit weniger für Verwirrung sorgen (Bowen et al., 2002; Harland, 2003). Ein weiteres Problem dieses Verfahrens ergibt sich bei der Personalselektion, da dessen Bearbeitung offenbar die intellektuellen Fähigkeiten stärker beansprucht als dies bei einem likert-skalierten Verfahren der Fall ist und somit neben Persönlichkeits- auch Intelligenzaspekte gemessen werden (*cognitive loading*; Christiansen et al., 2005; Vasilopoulos, Cucina, Dyomina, Morewitz & Reilly, 2006).

Oben dargestellte Ausführungen geben berechtigten Grund zur Annahme, dass ein auf der Basis des Wertequadrates entwickelter Persönlichkeits-Fragebogen mit Forced-Choice-Antwortformat resistenter gegenüber Faking ist als ein likert-skalierter Fragebogen. Dies liegt vor allem darin begründet, dass beim

Wertequadrat das Je-mehr-desto-besser-Prinzip nicht gilt und dass die beiden Tugenden prinzipiell gleichwertig sind, wobei es vom Situationskontext abhängt, welche Verhaltensweise angemessener sein könnte. Somit liessen sich mit dem Einsatz des Wertequadrates bei der Entwicklung von Persönlichkeitsskalen für den Einsatz in der Personalselektion drei Ziele gleichzeitig erreichen:

1. Die Logik des Wertequadrates gibt ein Konstruktionsraster vor, welches hervorragend für die Formulierung von Alternativ-Verhaltensweisen zu einem Item-Stamm geeignet ist.
2. Ein auf der Basis des Wertequadrates entwickelter Persönlichkeits-Fragebogen führt primär zu einer Beschreibung der Persönlichkeit und nicht zu einer Bewertung. Dies erleichtert es dem Personalverantwortlichen, dem Bewerber in jedem Fall eine positive, selbstwertschützende Rückmeldung zu geben, welche auch Entwicklungshinweise enthält.
3. Das in einem wertequadartbasierten Persönlichkeits-Fragebogen eingesetzte Forced-Choice-Format könnte die in Selektionssituationen auftretende Faking-Tendenz verringern.

4.6 Literaturverzeichnis

Aristoteles (1972). *Nikomachische Ethik.* Hamburg: F. Meiner.

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational and Organizational Psychology, 69*, 49–56.

Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology, 81*, 261–272.

Bartram, D. (1996). The relationship between ipsatized and normative measures of personality. *Journal of Occupational and Organizational Psychology, 69*, 25–39.

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment, 15*, 263–272.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on

personality measures. *International Journal of Selection and Assessment, 14*, 317–335.

Birkhan, G. (1998). Das Einzel-Assessment: Anatomie eines der wichtigsten Tage im Leben des Managers Herr Y. In M. Kleinmann & B. Strauss (Hrsg.), *Potentialfeststellung und Personalentwicklung* (S. 151–172). Göttingen: Verlag für Angewandte Psychologie.

Birkhan, G. (2007). Das unipolare und das bipolare Eigenschaftsmodell in Diagnostik und Beratung. In F. Westermann (Hrsg.), *Entwicklungsquadrat. Theoretische Fundierung und praktische Anwendungen* (S. 21–29). Göttingen: Hogrefe.

Birkhan, G. & Reitzig, G. (2007). Das Wertequadrat in der persönlichen Standortbestimmung. In F. Westermann (Hrsg.), *Entwicklungsquadrat. Theoretische Fundierung und praktische Anwendungen* (S. 83–93). Göttingen: Hogrefe.

Blickle, G. (1993). Ist Führen immer ein auswegloses Unterfangen? *Zeitschrift für Personalforschung, 7*, 202–415.

Bowen, C.-C., Martin, B. A., & Hunt, S. T. (2002). A comparison of ipsative and normative approaches for ability to control faking in personality questionnaires. *The International Journal of Organizational Analysis, 10*, 240–259.

Briefs, D. (2007). Gutachtenerstellung als Entscheidungshilfe bei der Führungskräfte-Auswahl. In F. Westermann (Hrsg.), *Entwicklungsquadrat. Theoretische Fundierung und praktische Anwendungen* (S. 69–82). Göttingen: Hogrefe.

Christiansen, N. D. (2008, October). *Faking matters: Effects on the usefulness of personality tests.* Papier vorgestellt am Gästekolloquium des Psychologischen Institutes der Universität Zürich, Fachrichtung Sozial- und Wirtschaftspsychologie.

Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance, 18*, 267–307.

Christiansen, N. D., Goffin, R. D., Johnston, N. G., & Rothstein, M. G. (1994). Correcting the 16PF for faking: Effects on criterion-related validity and individual hiring decisions. *Personnel Psychology, 47*, 847–860.

Converse, P. D., Oswald, F. L., Imus, A., Hedricks, C., Roy, R., & Butera, H. (2008). Comparing personality test formats and warnings: Effects on

criterion-related validity and test-taker reactions. *International Journal of Selection and Assessment, 16*, 155–169.

Converse, P. D., Peterson, M. H., & Griffith, R. L. (2009). Faking on personality measures: Implications for selection involving multiple predictors. *International Journal of Selection and Assessment, 17*, 47–60.

Cornwell, J. M., & Dunlap, W. P. (1994). On the questionable soundness of factoring ipsative data: A response to Saville & Willson (1991). *Journal of Occupational and Organizational Psychology, 67*, 89–100.

D'Heureuse, L. (1951). Gedanken zum Wertequadrat. *Psyche, 5 (2),* 117–119.

Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of applicant faking using randomized response technique. *Human Performance, 16*, 81–106.

Dunnette, M. D., McCartney, J., Carlson, H. C., & Kirchner, W. K. (1962). A study of faking behavior on a forced-choice self-description checklist. *Personnel Psychology, 15*, 13–24.

Dwight, S. A., & Donovan, J. J. (2003). Do warnings not to fake reduce faking? *Human Performance, 16*, 1–23.

Eberle, W. (2007). Das KEH-System und sein Interviewverfahren bei Personaleinstellungen. In F. Westermann (Hrsg.), *Entwicklungsquadrat. Theoretische Fundierung und praktische Anwendungen* (S. 45–57). Göttingen: Hogrefe.

Eberle, W. & Hartwich, E. (1995). *Brennpunkt Führungspotential. Persönlichkeitseinschätzung als unternehmerische Aufgabe.* Frankfurt am Main: Frankfurter Allgemeine Zeitung.

Edwards, A. L. (1954). *Edwards Personal Preference Schedule.* New York, NY: Psychological Corporation.

Ellingson, J. E., Sackett, P. R., & Hough, L. H. (1999). Social desirability corrections in personality measurement: Issues of applicant comparison and construct validity. *Journal of Applied Psychology, 84*, 155–166.

Erb, E. (1992). *Die Konstruktion menschlichen Denkens zwischen Dogmatismus als kurzschlüssiger Polarisierung und polarer Integration als Entwicklungsziel* (Bericht aus dem Psychologischen Institut, Nr. 75). Heidelberg: Universität Heidelberg, Psychologisches Institut.

Fleishman, E. A. (1953). The description of supervisory behavior. *Personnel*

Psychology, 37, 1–6.

Gloor, A. (1993). *Die AC-Methode. Assessment Center. Führungskräfte beurteilen und fördern.* Zürich: Orell Füssli.

Gloor, A. (2007a). Die Verheiratung des Big-Five-Konzeptes mit dem Wertequadrat-Modell – ein Entwurf. In F. Westermann (Hrsg.), *Entwicklungsquadrat. Theoretische Fundierung und praktische Anwendungen* (S. 31–44). Göttingen: Hogrefe.

Gloor, A. (2007b). Das Wertequadrat als „Feedback-Facilitator" und seine Anwendung in der Führungskräfte-Entwicklung. In F. Westermann (Hrsg.), *Entwicklungsquadrat. Theoretische Fundierung und praktische Anwendungen* (S. 95–125). Göttingen: Hogrefe.

Gloor, A. (2007c). *Das Werte- und Entwicklungsquadrat.* Unterlagen zur Lehrveranstaltung „Das Wertequadrat als Denkmuster im HRM". Heruntergeladen am 29. Juni 2010 von ftp://ftp.unizh.ch/hrm/03_studium/veranstaltungen/hrm_2/Folien_HRMII_SS07/2_Gastreferat_Armin_Gloor.pdf

Goffin, R. D., & Woods, D. M. (1995). Using personality testing for personnel selection: Faking and test-taking inductions. *International Journal of Selection and Assessment, 3*, 227–236.

Graham, W. R. (1958). An experimental comparison of methods to control faking of inventories. *Educational and Psychological Measurement, 18*, 387–401.

Griffith, R. L., Chmielowski, T., & Yoshita, Y. (2007). Do applicants fake? An examination of the frequency of applicant faking behavior. *Personnel Review, 36*, 341–355.

Griffith, R. L., & Peterson, M. H. (Eds.). (2006). *A closer examination of applicant faking behavior.* Greenwich, CT: IAP.

Griffith, R. L., & Peterson, M. H. (2008). The failure of social desirability measures to capture applicant faking behavior. *Industrial and Organizational Psychology, 1*, 308–311.

Groeben, N. (1981). Zielideen einer utopistisch-moralischen Psychologie. *Zeitschrift für Sozialpsychologie, 12*, 104–133.

Hakstian, A. R., & Ng, E.-L. (2005). Employment-related motivational distortion: It's nature, measurement, and reduction. *Educational and Psychological Measurement, 65*, 405–441.

Halpin A. W., & Winer, B. J. (1957). A factorial study of the leader behavior

descriptions. In R. M. Stogdill & A. E. Coons (Eds.), *Leader behavior: Its description and measurement* (pp. 39–51). Columbus, OH: Bureau of Business Research, Ohio State University.

Harland, L. K. (2003). Using personality tests in leadership development: Test format effects and the mitigating impact of explanations and feedback. *Human Resource Development Quarterly, 14*, 285–301.

Heggestad, E. D., Morrison, M., Reeve, C. L., & McCloy, R. A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology, 91*, 9–24.

Helwig, P. (1936). *Charakterologie*. Leipzig: Teubner.

Helwig, P. (1948). Das Wertequadrat. *Psyche, 2 (1)*, 121–127.

Helwig, P. (1951). *Charakterologie* (2. veränd. Aufl.). Leipzig: Teubner.

Helwig, P. (1967). *Charakterologie*. Freiburg: Herder.

Hemphill, J. K., & Coons, A. E. (1957). Development of the leader behavior description questionnaire. In R. M. Stogdill & A. E. Coons (Eds.), *Leader behavior: Its description and measurement* (pp. 6–38). Columbus, OH: Bureau of Business Research, Ohio State University.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74*, 167–184.

Hirsh, J. B., & Peterson, J. B. (2008). Predicting creativity and academic success with a ‘‘Fake-Proof” measure of the Big Five. *Journal of Research in Personality, 42*, 1323–1333.

Hogan, J., Barrett, P., & Hogan, R. (2007). Personality measurement, faking, and employment selection. *Journal of Applied Psychology, 92*, 1270–1285.

Holden, R. R., & Book, A. S. (2009). Using hybrid Rasch-latent class modeling to improve the detection of fakers on a personality inventory. *Personality and Individual Differences, 47*, 185–190.

Holden, R. R., & Hibbs, N. (1995). Incremental validity of response latencies for detecting fakers on a personality test. *Journal of Research in Personality, 29*, 362–372.

Holden, R. R., Wood, L. L., & Tomashewski, L. (2001). Do response time limitations counteract the effect of faking on personality inventory validity? *Journal of Personality and Social Psychology, 81*, 160–169.

Hough, L. M. (1998). Effects of intentional distortion in personality measurement

and evaluation of suggested palliatives. *Human Performance, 11*, 209–244.

Hough, L. M., Eaton, N. K., Dunnette, M. D., Kamp, J. D., & McCloy, R. A. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. *Journal of Applied Psychology, 75,* 581–595.

Jackson, D. N., Wroblewski, V. R., & Ashton, M. C. (2000). The impact of faking on employment tests: Does forced choice offer a solution? *Human Performance, 13*, 371–388.

Jaspers, K. (1919). *Psychologie der Weltanschauungen*. Berlin: Julius Springer.

Johnson, C. E., Wood, R., & Blinkhorn, S. F. (1988). Spuriouser and spuriouser: The use of ipsative personality tests. *Journal of Occupational Psychology, 61*, 153–162.

Kahn, R. L., & Katz, D. (1953). Leadership practices in relation to productivity and morale. In D. Cartwright & A. Zander (Eds.), *Group dynamics* (pp. 554–571). New York, NY: Harper & Row.

Klages, L. (1917). *Handschrift und Charakter. Gemeinverständlicher Abriss der graphologischen Technik*. Leipzig: Johann Ambrosius Barth.

Kronshage, U. (1995). Kommunikationspsychologische Hilfestellungen bei Wertekonflikten. In V. Heyse & H. Metzler (Hrsg.), *Die Veränderung managen, das Management verändern: Personal- und Organisationsentwicklung im Übergang zu neuen betrieblichen Strukturen – Trainingskonzepte zur Erhöhung von Kompetenzen* (S. 347–358). Münster: Waxmann.

Lammers, F. & Frankenfeld, V. (1999). Effekte gezielter Antwortstrategien bei einem Persönlichkeitsfragebogen mit "forced-choice"-Format. *Diagnostica, 45*, 65–68.

Li, A., & Bagger, J. (2006). Using the BIDR to distinguish the effects of impression management and self-deception on the criterion validity of personality measures: A meta-analysis. *International Journal of Selection and Assessment, 14*, 131–141.

McFarland, L. A. (2003). Warning against faking on a personality test: Effects on applicant reactions and personality test scores. *International Journal of Selection and Assessment, 11*, 265–276.

Meade, A. W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational and*

Organizational Psychology, 77, 531–552.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007a). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology, 60*, 683–729.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K., & Schmitt, N. (2007b). Are we getting fooled again? Coming to terms with limitations in the use of personality tests for personnel selection. *Personnel Psychology, 60*, 1029–1049.

Mueller-Hanson, R., Heggestad, E. D., & Thornton, G. C., III. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology, 88*, 348–355.

Müller, W. H. & Enskat, E. (1993). *Graphologische Diagnostik. Ihre Grundlagen, Möglichkeiten und Grenzen* (4. korr. u. erg. Aufl.). Bern: Huber.

Neuberger, O. (1983). Führen als widersprüchliches Handeln. *Psychologie und Praxis. Zeitschrift für Arbeits- und Organisationspsychologie, 27,* 22–32.

Neuberger, O. (1995). *Führen und geführt werden* (5. Aufl.). Stuttgart: Enke.

Neuberger, O. (2002). *Führen und führen lassen: Ansätze, Ergebnisse und Kritik der Führungsforschung* (6., völlig neu bearb. und erw. Aufl.). Stuttgart: Lucius und Lucius.

Ofman, D. (1992). *Bezieling en kwaliteit in organisaties.* Utrecht: Servire.

Ofman, D. (2005). *Qualität und Inspiration. Zugangswege zur Kreativität.* Berlin: WiKu.

Ones, D. S., Dilchert, S., Viswesvaran, C., & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology, 60*, 995–1027.

Ones, D. S, Viswesvaran, C., & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81*, 660–679.

Peseschkian, N. (1977). *Positive Psychotherapie – Theorie und Praxis einer neuen Methode.* Frankfurt: Fischer.

Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology, 78*, 582–593.

Robson, S. M., Jones, A., & Abraham, J. (2008). Personality, faking, and convergent validity: A warning concerning warning statements. *Human Performance, 21*, 89–106.

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology, 83*, 634–644.

Rothstein, M. G., & Goffin, R. D. (2000). The assessment of personality constructs in industrial–organizational psychology. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 215–248). Norwell, MA: Kluwer Academic.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review, 16*, 155–180.

Schmitt, N., & Oswald, F. L. (2006). The impact of corrections for faking on the validity of noncognitive measures in selection settings. *Journal of Applied Psychology, 91*, 613–621.

Schulz von Thun, F. (1989). *Miteinander Reden 2. Stile, Werte und Persönlichkeitsentwicklung.* Reinbek bei Hamburg: Rowohlt.

Schulz von Thun, F. (2000). *Miteinander Reden: Kommunikationspsychologie für Führungskräfte.* Reinbek bei Hamburg: Rowohlt.

Scott, W. A. (1963). Social desirability and individual conceptions of the desirable. *Journal of Abnormal and Social Psychology, 67*, 574–585.

Tett, R. P., & Christiansen, N. D. (2007). Personality tests at the crossroads: A response to Morgeson, Campion, Dipboye, Hollenbeck, Murphy, and Schmitt. *Personnel Psychology, 60*, 967–993.

Travers, R. M. W. (1951). A critical review of the validity and rationale of the forced-choice technique. *Psychological Bulletin, 48*, 62–70.

Vasilopoulos, N. L., Cucina, J. M., Dyomina, N. V., Morewitz, C. L., & Reilly, R. R. (2006). Forced-choice personality tests: A measure of personality and cognitive ability? *Human Performance, 19*, 175–199.

Vasilopoulos, N. L., Cucina, J. M., & McElreath, J. M. (2005). Do warnings of response verification moderate the relationship between personality and cognitive ability? *Journal of Applied Psychology, 90*, 306–322.

Vasilopoulos, N. L., Reilly, R. R., & Leaman, J. A. (2000). The influence of job

familiarity and impression management on self-report measure scale scores and response latencies. *Journal of Applied Psychology, 85*, 50–64.

Viswesvaran, C., & Ones, D. S. (1999). Meta-analysis of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement, 59*, 197–210.

Westermann, F. (Hrsg). (2007a). *Entwicklungsquadrat. Theoretische Fundierung und praktische Anwendungen.* Göttingen: Hogrefe.

Westermann, F. (2007b). Wer einen Schlüssel hat, der Türen öffnet, braucht nicht durch die Wand zu gehen! Das Entwicklungsquadrat – eine Einführung. In F. Westermann (Hrsg.), *Entwicklungsquadrat. Theoretische Fundierung und praktische Anwendungen* (S. 9–19). Göttingen: Hogrefe.

Wieser, R. (1960). *Mensch und Leistung in der Handschrift. Aus der Praxis der Betriebsgraphologie*. München: Ernst Reinhardt.

Zickar, M. J., Gibby, R. E., & Robie, C. (2004). Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. *Organizational Research Methods, 7*, 168–190.

5. Akzeptanz von Testverfahren[1]

5.1 Einführung in die Problematik der Durchführung psychologischer Testverfahren und die Erforschung deren Akzeptanz

> Die aus klinischen Tests entwickelten und zur Personalauslese umfunktionierten projektiven Verfahren nutzen oft auf raffinierte Weise die Arglosigkeit der Testpersonen aus und trachten danach, den Bewerber oder Mitarbeiter umfassend seelisch zu durchleuchten. Hier kann tatsächlich der Mensch zum Untersuchungsobjekt herabgewürdigt werden, zu einer Sache, die analysiert, registriert und katalogisiert werden kann, eben zugänglich wird für eine umfassende, zum Teil menschenunwürdige Bestandesaufnahme. Man muss sich deshalb wirklich wundern, dass viele Firmen von ihren Bewerbern und Mitarbeitern eine solche die Selbstachtung zerstörende und die Persönlichkeitssphäre zutiefst verletzende Selbstäusserung verlangen. (Schmid, 1971, S. 6)

Psychologische Testverfahren kamen im deutschen Sprachraum zu Beginn der siebziger Jahre stark in Verruf, nicht zuletzt auch deshalb, weil Psychologen diese trotz völliger Unangemessenheit im Personalbereich einsetzten, wie eine Auswahl aus der 1971 von Schmid veröffentlichten Liste zeigt: Wartegg-Zeichen-Test, Rosenzweig Picture Frustration Test, Szondi-Test, Thematic Apperception Test, Rorschach-Test, Farbpyramiden-Test, Lüscher-Test, Minnesota-Persönlichkeitsfragebogen. Es handelt sich hierbei durchwegs um klinische und/oder projektive Verfahren, welche keinerlei Bezug zur Arbeitswelt haben und für den Bewerber völlig intransparent sind.

Dies prangerte 1974 auch von Paczensky in ihren Buch „Testknacker" an, dessen erstes Kapitel die Überschrift „Der Tester und sein Opfer" trägt. Ziel dieses Buches war es, durch gezielte Provokation der mit der Personalselektion beauftragten Testpsychologen eine öffentliche Diskussion auszulösen, um aus Bewerbern – von ihr als menschenunwürdig behandelte Versuchspersonen bezeichnet – gleichberechtigte Partner zu machen. Auch Grubitzsch (1978) wies im viel beachteten Buch „Testtheorie – Testpraxis" (Grubitzsch & Rexilius, 1978) unter anderem auf die Situation der Bewerber im Selektionsprozess hin: Diese

[1] Diesem Kapitel liegt ein Beitrag in einem Sammelband zu psychologischen Aspekten im Recht der Personalführung zu Grunde (Boss, 2005), welcher hierzu aktualisiert, überarbeitet und stark erweitert wurde.

erlebten die Testverfahren als unbestechliche, wissenschaftliche Instrumente, welche ihre berufliche Zukunft massgeblich beeinflussen können und so bedrohlich und angsterzeugend wirken.

Diese Kritikpunkte waren jedoch nicht neu: Sieber (1969) berichtet, dass Testkritiker schon in den fünfziger Jahren anlässlich psychologischer Kongresse darauf hingewiesen haben, dass psychologische Tests den Menschen zum Testobjekt degradieren, diese in den sechziger Jahren jedoch verstummt seien. Zu den kritischen Stimmen aus den Reihen der Diagnostiker gesellten sich in den siebziger Jahren zusätzlich solche, welche die Gefahr des Einsatzes psychologischer Testverfahren aus einer gesellschaftspolitischen Perspektive beurteilten: „Der Begriff Selektion findet sein gesellschaftliches Anwendungsfeld dort, wo Herrschaftsinteresse darauf besteht, klassenmässige Trennung, Unterscheidung und Gruppierung aufrecht zu erhalten, und wo gesellschaftliches Oben und Unten mit biologischem Hoch- und Minderwert gleichgesetzt werden" (Hehlen, 1978, S. 98). Diese Zeit der massiven Auflehnung gegen die damalige Praxis der psychologischen Diagnostik führte im deutschen Sprachraum zu engagierten fachlichen Auseinandersetzungen (z. B. Pawlik, 1976; Pulver, Lang & Schmid, 1978; Schweizerische Gesellschaft für Psychologie, 1975, 1976; Spörli, 1978; Triebe & Ulich, 1977) und wurde als Krise in der psychologischen Diagnostik bezeichnet (z. B. Pulver, 1975; Westmeyer, 2004). Dabei nahmen die betroffenen Psychodiagnostiker eine eher zögerliche Abwehrhaltung ein und entwickelten kaum Ideen, wie man dieser Krise durch eine Weiterentwicklung der Prozesse und Methoden begegnen könnte. Immerhin führten diese Diskussionen dazu, dass die Berufsverbände der Psychodiagnostiker die Forderung nach einer humanen und fairen Testdurchführung in ihren Richtlinien aufgenommen haben (z. B. DIN 33430 (DIN, 2002); Standards für pädagogisches und psychologisches Testen; (Häcker, Leutner & Amelang, 1998); siehe auch Rauchfleisch, 1982).

Schuler und Stehle stellten sich der Herausforderung und unternahmen 1983 einen Versuch zur Rehabilitierung der psychologischen Eignungsdiagnostik, indem sie in einer wissenschaftlichen Publikation den ersten konkreten Lösungsansatz zum oft berichteten Unbehagen der Testkandidaten präsentierten. Sie postulierten in ihrem Konzept zur sozialen Validität – welches ich in Kapitel 5.3 ausführlich darstelle – vier Aspekte respektive Komponenten, welche als Richtschnur für die Weiterentwicklung eignungsdiagnostischer Verfahren hinsichtlich Bewerberfreundlichkeit dienen sollen: Information, Partizipation, Transparenz und Urteilskommunikation. Schuler (1993a) gelang es, das Konzept über den deutschen Sprachraum hinweg bekannt zu machen, wodurch er einen wichtigen Beitrag für die weitere Erforschung der Reaktionen von Bewerbern auf Selektionsprozesse lieferte. Einzelne Aspekte davon übernahm eine Expertengruppe

schliesslich in die DIN 33430 „Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen" (2002; siehe auch Boss, 2005; Hornke & Winterfeld, 2004; Kanning, 2004; Westhoff, 2006; Westhoff et al., 2004).

Anders verlief der Prozess der Bewusstmachung der Bewerberperspektive in den Vereinigten Staaten von Amerika: Dort befassten sich zu Beginn der sechziger Jahre Untersuchungsausschüsse des Kongresses mit der Frage der Verletzung von Bürgerrechten durch den Einsatz von Selektionsverfahren, hauptsächlich Persönlichkeitstests im Rahmen der Auswahl von Beamten. 1964 verabschiedete der Kongress den *Civil Rights Act*, dessen Kapitel VII vorschreibt, dass Arbeitgeber (und auch die von ihnen eingesetzten Testverfahren) Arbeitnehmende nicht auf Grund der Rasse, Hautfarbe, Religion, Geschlecht oder Herkunft bei der Einstellung, der Beförderung oder der Entlassung diskriminieren dürfen. Der *Equal Employment Opportunity Act* von 1972 ergänzte das Kapitel VII unter anderem um die Aspekte Alter und Behinderung, die Revision aus dem Jahre 1995 führt insgesamt schon 13 Aspekte auf. Über die Diskriminierung hinausgehende Aspekte der Bewerberperspektive sind in diesen Gesetzen, wie auch in den von der *Equal Employment Opportunity Commission* erarbeiteten *Uniform Guidelines on Employee Selection Procedures* (1966, 1978), jedoch nicht enthalten. (Für weitere Ausführungen zu den gesetzlichen Bestimmungen der Personalselektion in den USA siehe z. B. Cascio & Aguinis, 2005.) Der Amerikanische Kongress untersuchte 1965 den Gebrauch und den Missbrauch psychologischer Testverfahren und psychiatrischer Abklärungen, indem er deren Validität, Reliabilität und Angemessenheit im Hinblick auf den Einsatz in der Personalselektion überprüfte. Dieser Kontroverse, welche das Selbstverständnis diagnostisch tätiger Psychologen angriff und diese so in Alarmbereitschaft versetzte, hat die *American Psychological Association* die November-Ausgabe 1965 ihres Fachorgans – der *American Psychologist* – gewidmet. Den Stellenwert der in dieser Zeit geführten Diskussion belegt auch der Beitrag von Remmers, Leidy, Starry, Shuman und Tesser von 1966 mit dem Titel „High school students' attitudes on two controversial issues: War in Southeast Asia and the use of personality and ability tests". Viele US-amerikanische Forscher beschäftigten sich auf Grund dieser Gesetzesartikel mit der Frage der Fairness respektive dem *adverse impact* – der unbeabsichtigten Diskriminierung bei der Durchführung von Selektionsverfahren –, so dass diese Thematik über lange Zeit eine dominierende Stellung in der Literatur zur Bewerberperspektive in Selektionsprozessen einnahm (erste Übersichtsdarstellung: Kirkpatrick, Ewen, Barrett & Katzell, 1968; aktuelle Übersichtsdarstellungen: Hough, Oswald & Ployhart, 2001; Ployhart & Holtz, 2008).

Erst 1992 erweiterten Schmitt und Gilliland das Konzept der Fairness um den Aspekt der Bewerberreaktion auf Selektionsentscheide und unterschieden so

zwischen den psychometrischen Aspekten und der Wahrnehmung des Bewerbers einer fairen Selektion. Im gleichen Jahr erschien der erste Sammelband von Saunders zu dieser Thematik (*New approaches to employee management: Fairness in employee selection*). Gilliland (1993) übertrug die aus der Forschung zur *Organizational Justice* gewonnenen Erkenntnisse (für Übersichtsdarstellungen siehe Colquitt, Colon, Wesson, Porter & Ng, 2001; Greenberg, 1990) auf die wahrgenommene Fairness von Selektionsverfahren, entwickelte ein Modell der Bewerberreaktionen und formulierte zehn Regeln für faire Selektionsprozesse: Tätigkeitsbezug, Möglichkeit zur Selbstdarstellung, Möglichkeit zur Wiedererwägung, Vergleichbarkeit der Durchführung, Ergebnisrückmeldung, Information zum Auswahlverfahren, Aufrichtigkeit, respektvolle Behandlung, Zweiweg-Kommunikation und Angemessenheit der Fragen. Mit seinem Beitrag erweiterte Gilliland den Fokus der Fairnessforschung und inspirierte viele Forscher, sich vertiefter mit dieser vielfältigen Thematik zu befassen (z. B. Anderson & Witvliet, 2008; Bauer, Maertz, Dolen & Campion, 1998; Bauer, Truxillo, Sanchez, Craig, Ferrara & Campion, 2001; Bell, Ryan & Wiechmann, 2004; Bell, Wiechmann & Ryan, 2006; Burns, Siers & Christiansen, 2008; Chan, 1997; Chan, Schmitt, Jennings, Clause & Delbridge, 1998; Converse, Oswald, Imus, Hedricks, Roy & Butera, 2008; Cropanzano & Konovsky, 1996; Donovan, Drasgow & Munson, 1998; Gilliland, Groth, Baker, Dew, Polly & Langdon, 2001; Harold & Ployhart, 2008; Hausknecht, Day & Thomas, 2004; Horvath, Ryan & Stierwalt, 2000; Kravitz, Stinson & Chavez, 1996; Lievens, De Corte & Brysse, 2003; Ployhart & Ryan, 1997, 1998; Ployhart, Ryan & Bennett, 1999; Rolland & Steiner, 2007; Sanchez, Truxillo & Bauer, 2000; Schmit & Ryan, 1997; Truxillo, Bauer, Campion & Paronto, 2002; Truxillo, Steiner & Gilliland, 2004). Auf das Modell von Gilliland gehe ich ausführlich in Kapitel 5.4 ein.

Auf Grund der Analyse der zwischen 1985 und 1999 erschienenen Artikel zu Bewerberreaktionen stellten Ryan und Ployhart (2000) ein integratives Modell auf, welches Hausknecht et al. (2004) erweiterten und differenzierten. Darin unterscheiden sie zwischen Ausgangsvariablen (Merkmale der Person, Merkmale des Selektionsprozesses, Merkmale der Stelle, organisationaler Kontext), Wahrnehmungen des Bewerbers während des Selektionsprozesses, Outcome-Variablen (Leistung im Selektionsverfahren, Selbstwahrnehmungen, Einstellungen und Verhalten gegenüber der Organisation, Arbeitsverhalten) und einer Reihe von Moderatorvariablen. In Kapitel 5.5 stelle ich dieses Modell dar.

Die Erforschung der Reaktionen von Bewerbern auf den Selektionsprozess in den letzten vierzig Jahre hat dazu geführt, dass heute zu vielen Themenbereichen, wie zum Beispiel der Fairness im Sinne der Gleichbehandlung, der Akzeptanz unterschiedlicher Testverfahren oder zu den Auswirkungen der

wahrgenommenen Fairness, für die Praxis wichtige Erkenntnisse vorliegen. Zudem entwickelten mehrere Forschergruppen zum Teil umfassende Modelle zur Bewerberreaktion, mit welchen sie die komplexen Prozesse der Wahrnehmung und Verarbeitung von Selektionsprozessen durch die Bewerber und die dadurch entstehenden Einstellungen und Verhaltensweisen zu erklären versuchen. Nachdem Anderson 2004 in seinem Editorial zu einem Special Issue des *International Journal of Selection and Assessment* bezüglich der Bewerberperspektive in Selektionsprozessen noch von „the dark side of the moon" (S.1) sprach, stellte er bereits fünf Jahre später fest, „that the dark side has been considerably enlightened" (Hülsheger & Anderson, 2009, S. 335), was er auf die in der Zwischenzeit blühende Forschungstätigkeit im Bereich der Bewerberperspektive und der Bewerberreaktionen zurückführt.

Da die Forschung zu den Bewerberreaktionen auf Testverfahren hauptsächlich durch die Modelle von Schuler und Gilliland beeinflusst wurden (Derous, Born & De Witte, 2004) gehe ich – nach der Darstellung der Auswirkungen der Akzeptanz von Selektionsprozessen auf die durchführende Organisation– vertieft auf diese beiden ein. In Kapitel 5.5 stelle ich ergänzend dazu noch das Modell von Hausknecht et al. und die Studie zur Struktur der Bewerberreaktionen im Militär von Schreurs vor. Im Anschluss daran beschreibe ich im historischen Überblick Skalen zur Erfassung der Akzeptanz von Testverfahren, des Fairnessempfindens in Selektionssituationen und der Bewerberreaktionen auf Selektionsprozesse. Den Abschluss dieses Kapitels bilden Ausführungen zu Unterschieden in der Akzeptanz der am häufigsten in der Personalselektion eingesetzten Testverfahren.

5.2 Einfluss der wahrgenommenen Fairness eines Selektionsverfahrens auf die Einstellungen der Bewerber gegenüber der Organisation

In diesem Kapitel gehe ich der Frage nach, wieso es für die Organisation wichtig ist, dass die Bewerber ihren Selektionsprozess gut akzeptieren. Auch wenn – wie nachfolgendes Zitat zeigt – schon lange bekannt ist, dass ein von den Bewerbern schlecht akzeptiertes Selektionsverfahren Auswirkungen auf deren Einstellungen gegenüber der durchführenden Organisation haben kann, befassen sich Forscher erst seit weinigen Jahren intensiv mit dieser Thematik. Packard (1966, zit. nach Schmid, 1971, S. 17) schrieb dazu: „Jede Firma, die ihre zukünftigen Angestel-

lten einer derartig unwürdigen Überprüfung aussetzt, verdient von ihnen im Hinblick auf Loyalität, Einsatzbereitschaft und Aufrichtigkeit keine guten Leistungen und wird sie auch nicht bekommen." Schuler (1998, S. 193) nennt eine Reihe von Problemen, welche beim unreflektierten, laienhaften Einsatz von Testverfahren bei der Personalselektion auftreten können:

- Verwendung unzulänglicher Diagnosemethoden
- Vernachlässigung des Anforderungsbezuges
- Anwendung unnötig belastender Verfahren
- Eindringen in die Privatsphäre der Kandidaten
- Nötigung zur Exposition unerwünschten Verhaltens (z. B. in Gruppen)
- Mangelnde Vertraulichkeit der erhobenen Daten
- Missbrauch diagnostischer Daten
- Mangelnde Rücksichtnahme in der Urteilskommunikation
- Informationsverweigerung bezüglich der Ergebnisse
- Konflikt zwischen der Informationsverpflichtung gegenüber dem Auftraggeber und den Interessen der Kandidaten

In ihrem Modell zu Bewerberreaktionen führen Hausknecht et al. (2004, ausführliche Darstellung siehe Kapitel 5.5.1) eine ganze Reihe möglicher Auswirkungen von Selektionsverfahren auf, welche sie aufgeteilt haben in die objektiv erbrachte und die subjektiv eingeschätzte Leistung im Selektionsverfahren, die Selbstwahrnehmungen des Bewerbers (Selbstwirksamkeit, Selbstwertgefühl), die Arbeitseinstellungen und das Arbeitsverhalten des eingestellten Bewerbers (Arbeitszufriedenheit, organisationales Commitment, Arbeitsleistung, Organizational Citizenship Behaviors, Kündigungsabsichten resp. Kündigung) und die Einstellungen und das Verhalten des Bewerbers gegenüber der Organisation (Attraktivität der Organisation, Eingehen auf ein Angebot, Weiterempfehlung, Bewerbung und Wiederbewerbung, Retesting, Produkt-Kauf, Erheben einer Anklage, Rückzug der Bewerbung). Gut nachvollziehbar ist in diesem Zusammenhang, dass abgewiesene Bewerber ein Testverfahren oder den Selektionsprozess als deutlich weniger fair einstufen als aufgenommene (Kluger & Rothstein, 1993; Schleicher, Venkataramani, Morgeson & Campion, 2006) und die Organisation weniger attraktiv finden (Anseel & Lievens, 2009; Ryan & Ployhart, 2000). Bauer et al. (1998) sind auf Grund ihrer Studienergebnisse der Ansicht, dass das Abschneiden im Selektionsverfahren die Einstellung gegenüber der Organisation stärker beeinflusst, als die eingestufte Fairness der dabei eingesetzten Testverfahren. Die Forschungsergebnisse der letzten zwanzig Jahre zeigen jedoch auf,

dass negative Haltungen nicht nur bei abgewiesenen Bewerbern auftreten können, sondern auch bei den aufgenommenen.

Cropanzano und Wright (2003) unterteilen die Konsequenzen unfairer Behandlung während eines Personalselektionsverfahrens in die vier Bereiche Ablehnung eines Job-Angebotes, Erhöhung der Wahrscheinlichkeit eines Rechtsstreites, schlechte Meinung über die Organisation und Reduzierung der Arbeitsleistung. Nachfolgend werde ich diese vier Bereiche näher ausführen und Studien dazu vorstellen.

Ablehnung eines Job-Angebotes

Das beste Selektionsverfahren bringt schlussendlich nicht das erhoffte Resultat, wenn der Topkandidat das Job-Angebot ablehnt (Gilliland & Steiner, 1999). Dies führt im schlimmsten Fall dazu, dass der gesamte Selektionsprozess nochmals durchgeführt werden muss, vermindert jedoch auf jeden Fall die Nützlichkeit eines Test- oder Selektionsverfahrens (Boudreau & Rynes, 1985). Zum Beispiel konnten Crant und Bateman (1990) in Rollenspielen, Rosse, Ringer und Miller (1996, siehe auch Kluger & Rothstein, 1993) in experimentellen Studien und Stoffey, Millsap, Smither und Reilly (1991) in Feldstudien nachweisen, dass bei Bewerbern, welche die Testverfahren als unfair eingestuft haben, die Bereitschaft sinkt, ein Job-Angebot anzunehmen. Die Ergebnisse der Feldstudie von Madigan und Macan (2005) zeigen, dass sechs Aspekte der Fairness des Selektionsverfahrens mit der Gesamtbeurteilung der Fairness zusammenhängen, welche ihrerseits die Intention der Bewerber, die Organisation weiterzuempfehlen und ein Job-Angebot anzunehmen beeinflusst. In der Studie von Schmit und Ryan (1997) zogen gut 10% der Kandidaten auf die Position eines Polizei-Offiziers ihre Bewerbung zurück, weil sie das Gefühl hatten, ungerecht behandelt zu werden. In anderen Feldstudien zeigten sich jedoch nur geringe Effekte (Bauer et al., 1998; Macan, Avedon, Paese & Smith, 1994). Zudem scheint es so zu sein, dass vor allem diejenigen ein Verfahren als unfair empfinden, welche darin nicht so gut abschneiden (Ryan, Sacco, McFarland & Kriska, 2000), weshalb anzunehmen ist, dass der Schaden durch den Rückzug einer Bewerbung für die Unternehmung gering ausfallen wird. Es ist nahe liegend, dass auf Grund der noch nicht gefestigten Beziehung zwischen dem Bewerber und der Unternehmung das Erleben einer Ungerechtigkeit geringere Auswirkungen hat, als bei einem langjährigen Mitarbeiter (Brockner, Tyler & Cooper-Schneider, 1992; Ryan, Ployhart, Greguras & Schmit, 1997). Weiter spielen natürlich die Chancen auf dem Arbeitsmarkt des Bewerbers eine Rolle: Wenn er dringend auf eine Arbeitsstelle angewiesen ist und sich ihm wenige Möglichkeiten bieten, wird er der unfairen Behandlung

während des Selektionsprozesses wenig Beachtung schenken und den Job annehmen.

Wahrscheinlichkeit eines Rechtsstreites

Die Befürchtung, auf Grund eines als unfair empfundenen Testverfahrens in einen Rechtsstreit mit einem Bewerber verwickelt zu werden, ist in den USA – wie in Kapitel 5.1 dargestellt – als durchaus berechtigt einzustufen. So handeln die Autoren amerikanischer Handbücher zum Human Resource Management oder zur Personalselektion auch an prominenter Stelle ausführlich die entsprechenden Gesetze und die zu befolgenden Vorgehensweisen ab (z. B. Cascio & Aguinis, 2005; Cook, 2004; Gatewood, Feild & Barrick, 2008). Es müssen jedoch nicht einmal grobe Gesetzesverstösse sein, welche Bewerber vor Gericht ziehen lassen, in einigen Fällen reicht auch nur schon das Gefühl, unfair behandelt worden zu sein, für das Erheben einer Anklage (Bies & Tyler, 1993). Dies führt dazu, dass zum Beispiel Manager strukturierte Interviews bevorzugen, unter anderem weil sie davon ausgehen, dass im Falle einer Anklage diese Methode vor Gericht bestehen würde (Latham & Finnegan, 1993). In Europa hat dieses Thema (noch) eine untergeordnete Bedeutung, da die Rechtslage bezüglich Personalselektion deutlich weniger streng ist als in den USA (Boss, 2005) und die Arbeitnehmer den Gang zum Arbeitsgericht eher scheuen.

Schlechte Meinung über die Organisation

Es existieren zahlreiche Studien, welche belegen, dass der Einsatz als unfair empfundener Testverfahren die Organisation bei den Bewerbern in einem schlechten Licht erscheinen lässt (z. B. Bauer et al., 2001; Kluger & Rothstein, 1993; Macan et al., 1994; Schmitt & Gilliland, 1992; Stoffey et al., 1991; Truxillo & Bauer, 1999). Diese negative Einschätzung lässt sich auch bei eingestellten Bewerbern nachweisen, welche ein unfaires Selektionsverfahren durchlaufen haben, wie Gilliland (1994) in einer Experimentalstudie aufzeigen konnte. Smither, Reilly, Millsap, Pearlman und Stoffey (1993) fanden einen Zusammenhang zwischen der Augenscheinvalidität der eingesetzten Testverfahren und der Einstufung der Attraktivität der Organisation. Bauer et al. (1998) weisen darauf hin, dass zwar das Empfinden von Unfairness zu einer negativen Einschätzung der Organisation führen kann, dass das Bild, welches sich ein Bewerber von der Organisation macht, jedoch von zahlreichen weiteren Faktoren beeinflusst wird. Schreurs, Derous, Proost, Notelaers und De Witte (2008) untersuchten den Zusammenhang zwischen den Erwartungen von Bewerbern auf eine Stelle als

Berufssoldat zum Selektionsprozess (Wärme/Respekt, Möglichkeit, sein Potenzial zu zeigen, Schwierigkeit des Betrügens, unvoreingenommene Beurteilung und Feedback) mit Angaben zu bewerberspezifischen Attraktivitätsindikatoren (Attraktivität der Organisation, Absicht, die Stelle anzutreten, Ansehen der Organisation, Übereinstimmung zwischen der Person und der Organisation). Die berechneten Korrelationen liegen – mit Ausnahme bei der Variable Schwierigkeit des Betrügens (r = .17) – zwischen r = .30 und .36.

Reduzierung der Arbeitsleistung

Es gibt vereinzelte Hinweise darauf, dass Personen, welche sich während des Selektionsprozesses unfair behandelt fühlten, später am Arbeitsplatz eine geringere Leistung zeigen (z. B. Ambrose & Cropanzano, 2003; Gilliland, 1994; Konovsky & Cropanzano, 1991). Da die Leistung am Arbeitsplatz multikausal bedingt ist, ist es jedoch schwierig, die gefundenen Unterschiede eindeutig auf die Akzeptanzwahrnehmung während des Selektionsprozesses zurückzuführen.

Ryan und Ployhart (2000; siehe auch Rynes, 1993) kritisieren die Mehrzahl der im Zusammenhang mit den Auswirkungen der Bewerberwahrnehmungen durchgeführten Studien, weil sie deren Einstellungen vor dem Absolvieren des Selektionsprozesses nicht erfassen:

> Without assessing attitudinal measures (e.g., organizational attractiveness) and intentions prior to participating in the selection process, one is hard pressed to be able to definitively attribute a causal order. That is, applicant perceptions of the selection procedure may cause intentions and attitudes, or these intentions and attitudes may lead one to hold certain perceptions of the procedure. For example ... one may view an organization as attractive and this can cause one to see the selection process in a more favorable light. (S. 593)

Schreurs et al. (2008) zeigten mit ihren Daten auf, dass die Erwartungshaltung die Einschätzung des absolvierten Selektionsprozesses beeinflusst. „From a theoretical perspective, these results suggest that applicants' posttest perceptions are partly a function of their pretest beliefs, and that applicants tend to see what they expect to see" (Schreurs et al., 2008, S. 175). Somit ist es wichtig, dass mit dem Aufgebot der Bewerber umfassend über den bevorstehenden Selektionsprozess informiert wird, damit sich dieser ein realistisches Bild davon machen kann. Bauer et al. (1998) zeigten zudem auf, dass zwischen der Infor-

miertheit über die Testverfahren und der Attraktivität der Organisation und der Testfairness ein Zusammenhang besteht.

5.3 Das Konzept der sozialen Validität von Schuler und Stehle

Bis zur Publikation des Beitrages von Schuler und Stehle (1983) zur sozialen Validität fand das Erleben des Testkandidaten in der testdiagnostischen Situation als Forschungsthema im deutschsprachigen Raum keine Beachtung. In seinem Beitrag zur Personalauslese erwähnt zum Beispiel Jäger (1961) lediglich, dass es allgemeine sittliche und gesetzliche Normen und soziale Gepflogenheiten zu beachten gilt und der Bewerber als Partner und nicht als Diagnoseobjekt anzusehen ist, ohne dabei auf entsprechende Studien zu verweisen.

Wie in der Einleitung erwähnt, entwickelten Schuler und Stehle ihr Konzept als Antwort auf die massive Kritik an der Psychodiagnostik und der dabei verwendeten Verfahren in den siebziger Jahren des vergangenen Jahrhunderts. Sie identifizierten drei Ansatzpunkte für Verbesserungsmöglichkeiten: die eingesetzten Methoden und Entscheidungsprozesse (technisch-empirische Methoden), das Erleben und die Wirkung eignungsdiagnostischer Situationen und die soziale Situation. In ihrem Konzept der sozialen Validität beschrieben sie die vier Parameter Information, Partizipation, Transparenz und Urteilskommunikation, deren Nichteinhaltung sie als Gründe „des oft berichteten Unbehagens der Testkandidaten in der eignungsdiagnostischen Situation" (Schuler & Stehle, 1983, S. 34) identifizierten. Nachfolgend stelle ich diese vier Parameter ausführlich dar (siehe auch Schuler, 1990, 1993a, 1998; Schuler & Funke, 1987; Schuler & Stehle, 1983, 1985). Empirische Studien zu diesen und weiteren Aspekten der Bewerberreaktionen auf Personalauswahlverfahren führe ich aus Gründen der Übersichtlichkeit im Kapitel zum Modell von Gilliland (1993) auf.

Information

Nicht nur der Arbeitgeber muss sich ein umfassendes Bild über den zukünftigen Mitarbeiter machen, um eine gute Entscheidung treffen zu können, sondern auch der Bewerber sollte sich darüber informieren können, was ihn am Arbeitsplatz erwartet, um seinerseits entscheiden zu können, ob er die Stelle antreten möchte – Schuler (1990) spricht hierbei von der Selbstselektion. Neben einer detaillierten Beschreibung des Arbeitsinhaltes, der Aufgabenbereiche, der Anforderungen der

Tätigkeit und der Merkmale und Ziele der Organisation, sollte der Personalverantwortliche respektive der Linienvorgesetzte auch über sozialpsychologische Aspekte wie den Führungsstil, die Art und Weise der Zusammenarbeit und die Unternehmenskultur informieren. Ergänzen lassen sich diese Ausführungen durch den Einsatz von Testverfahren, welche in einem direkten Zusammenhang mit der zukünftigen Tätigkeit stehende Fähigkeiten und Fertigkeiten erfassen, wie zum Beispiel Arbeitsproben oder Fallstudien (Tachler, 1983, zitiert nach Schuler, 1990).

Ein derart ausgestaltetes Selektionsverfahren ermöglicht es dem Bewerber nicht nur, sich für oder gegen die konkrete Stelle zu entscheiden, sondern sich auch ein klareres Bild seiner arbeitsbezogenen Neigungen und somit seiner beruflichen Zukunft zu machen. Somit würde dieser Prozess Ähnlichkeiten mit einer Berufsberatungssituation haben und es könnte – würde dies von den Personalverantwortlichen konsequent umgesetzt – ein grosser Schritt in Richtung Freiwilligkeit der Teilnahme an einem Selektionsverfahren erreicht werden, wie es Stoll (1977) als Ausweg aus der Krise der Diagnostik vorgeschlagen hat. Dieser Aspekt der Selbstselektion, welcher auf die Einsicht des Bewerbers zurückzuführen ist, dass er nicht auf diese Stelle oder in diese Unternehmenskultur passt, ist bis heute jedoch wenig untersucht worden. Hingegen ist gut belegt, dass ein schlecht akzeptiertes Selektionsverfahren dazu führt, dass Bewerber sich zurückziehen, weil sie nicht für eine Unternehmung arbeiten wollen, welche schon den Bewerber gegenüber unfaires Verhalten zeigen (z. B. Bretz & Judge, 1998; Ryan et al., 2000; Schmit & Ryan, 1997).

Partizipation

Wie bei anderen organisationspsychologischen Veränderungen sollen die Diagnostiker auch bei der Entwicklung und Durchführung eignungsdiagnostischer Verfahren betroffene Gruppen einbeziehen. Gute Beispiele dazu sind die Sammlung des Ausgangsmaterials im Rahmen der Entwicklung von Situational Judgment Tests oder die Kontrolle der Verständlichkeit und Akzeptanz der Aussagen in einem Persönlichkeits-Fragebogen. Neben dem grossen Nutzen dieser Informationen für den Testentwickler fördert die Einsicht in den Aufbau der in der Eignungsdiagnostik eingesetzten Methoden und Verfahren zudem deren Akzeptanz. Herberger (1984, zitiert nach Schuler, 1990) zeigte auf, dass der Hinweis auf den Einbezug von Auszubildenden bei der Gestaltung des Auswahlprozesses die Zufriedenheit der Bewerber mit dem Selektionsprozess, dem Diagnostiker, dem organisatorischen Ablauf und den Entwicklungsmöglichkeiten erhöht.

Im weiteren Sinne versteht Schuler (1990) unter Partizipation, dass der Bewerber Kontrolle über die Situation ausüben und seine Stärken zum Einsatz bringen kann. Er erhält so den Eindruck, dass er bewusst und mit seinen ihm zur Verfügung stehenden Fähigkeiten den Selektionsentscheid beeinflussen kann. Schuler und Funke (1987) sind der Ansicht, dass Interviews gerade auf Grund der wahrgenommenen Beeinflussbarkeit bei den Bewerbern zu den beliebtesten eignungsdiagnostischen Verfahren zählen. Im Endeffekt wirkt sich die Partizipation auch dahingehend aus, dass sich der Bewerber Ernst genommen und respektiert fühlt.

Transparenz

Während des gesamten Selektionsverfahrens muss dem Bewerber klar sein, welche Rollen die im Selektionsprozess beteiligten Personen einnehmen, welche Fähigkeiten der Diagnostiker untersucht und welches Verhalten dieser von ihm erwartet. Die Rollenklärung als zentrales Element des Schutzes der Kandidaten psychologischer Eignungsabklärungen forderte Stoll schon 1977. Zudem müssen die eingesetzten diagnostischen Verfahren derart ausgestaltet sein, dass der Bewerber einen klaren Zusammenhang zur Feststellung der gefragten beruflichen Eignung erkennen kann – diese also augenscheinvalide sind – und so versteht, zu welchem Zweck diese eingesetzt werden. Der Einsatz von anforderungsbezogenen und nachvollziehbar gestalteten Aufgaben soll dem Bewerber zudem die Möglichkeit zur Selbstbeurteilung bieten. Als weitere Anforderung an ein transparentes Verfahren hat der Diagnostiker dem Bewerber die Auswertungskriterien und Beurteilungsmassstäbe, die Prinzipien des diagnostischen Schlusses und die Transformation der Daten in Urteile und Entscheide verständlich zu erläutern.

Urteilskommunikation

Auch bei der abschliessenden Rückmeldung der Ergebnisse soll der Diagnostiker transparent, offen, wahrhaftig, rücksichtsvoll, unterstützend und nachvollziehbar kommunizieren. Dabei achtet er auf das Selbstwertgefühl des Bewerbers und nimmt auf dessen Persönlichkeitssphäre Rücksicht, indem er Distanz einhält und nur arbeitsrelevante Merkmale bespricht. Ein gut akzeptiertes Feedback erreicht der Diagnostiker zudem dadurch, dass er nicht Eigenschaften schildert, sondern verhaltensbezogene Aussagen macht und hauptsächlich auf Stärken und Entwicklungsmöglichkeiten hinweist. Ziel soll sein, dass der Bewerber – auch bei einer Ablehnung – einen Profit aus dem Vorstellungsgespräch und der investierten Zeit ziehen kann, indem er sich und seine Fähigkeiten besser kennen gelernt

und Klarheit über seine berufliche Eignung erlangt hat. Wie ich weiter unten bei den Ausführungen zum Modell von Gilliland (1993) zeige, muss der Feedbackgeber jedoch bei abgewiesenen Bewerbern besondere Vorsicht walten lassen, um deren emotionale Befindlichkeit nicht zu stark zu beeinträchtigen.

Mit der Veröffentlichung des Konzeptes zur sozialen Validität haben Schuler und Stehle den Aspekt des Erlebens des Bewerbers in der diagnostischen Situation als neues Forschungsgebiet etabliert und einige empirische Arbeiten dazu angeregt, so auch solche zum Assessment Center (z. B. Harburger, 1992; Runge, 1996; Sichler, 1989), welches Schuler und Stehle als Beispiel für die Darstellung ihres Konzeptes benutzten. Zudem diente das Konzept Forschern dazu, die Akzeptanz verschiedener Verfahren anhand eines einheitlichen Massstabes zu beurteilen (z. B. Kersting, 1998; Smither et al., 1993). Bis Mitte der 90er Jahre lassen sich noch Arbeiten finden, welche einen direkten Bezug dazu herstellen. Dass das Interesse an diesem Konzept abgenommen hat, kann auf verschiedene Ursachen zurückgeführt werden:

Unklar ist, welches die theoretischen und/oder empirischen Grundlagen des Konzeptes sind. Schuler (2002, S. 108-109) schreibt später dazu: „Als die mutmasslich wichtigsten Parameter, die Auswahlsituationen zu sozial akzeptablen Situationen machen, werden angenommen: ..." Köchling (2000) bezeichnet das Konzept als „eine reine Auflistung einiger möglicher Determinanten der Bewertung bzw. Akzeptanz von Auswahlsituationen" (S. 23). Es ist anzunehmen, dass es den Autoren nicht primär darum ging, das Bewerberverhalten und –erleben erklären zu können, sondern Ansatzpunkte für die Verbesserung von Selektionsprozessen zu liefern und sie deshalb Aspekte aufgenommen haben, welche Kritiker und Befürworter anlässlich der Diskussion zur Krise in der Diagnostik ins Feld geführt haben. So lassen sich zum Beispiel bei von Paczensky (1974) kritische Bemerkungen zu allen vier Parametern finden (siehe auch Boss, 2005). Das Konzept liefert somit nur einen Ausschnitt aus der Gesamtheit möglicher Bewerberreaktionen und es fehlen ihm zugrunde liegende allgemeinpsychologische Theorien, welche Erklärungsansätze für Zusammenhänge zwischen den einzelnen Parametern liefern. Dies ist mit ein Grund, weshalb es sich beim Konzept von Schuler und Stehle nicht um ein Modell handelt, sondern letztendlich nur um eine Sammlung möglicher Einflussfaktoren. Die beiden Autoren haben nie dargelegt, wie die einzelnen Parameter untereinander verknüpft sind und wie sich diese schlussendlich zum Beispiel auf die Akzeptanz der Testverfahren, die Zufriedenheit mit dem Auswahlprozess oder die Akzeptanz des Entscheides auswirken. Somit ist ihr Konzept in seiner Gesamtheit auch nicht

empirisch überprüfbar und es existieren heute lediglich Studien zur Angemessenheit der als unabhängig voneinander aufgefassten vier Parameter (z. B. Fruhner, Schuler, Funke & Moser, 1991; Schuler, 1993a. Siehe z. B. auch Chan, Schmitt, DeShon, Clause & Delbridge, 1997; Derous & De Witte, 2001; Macan et al., 1994; Schinkel, van Dierendonck & Anderson, 2004; Schreurs, 2007; Smither et al., 1993). Trotzdem wurde in den letzten zehn Jahren so viel wie noch nie zuvor zu den Reaktionen von Bewerbern auf Selektionsverfahren geforscht und publiziert. Diese Arbeiten beziehen sich jedoch häufig auf das Modell von Gilliland (1993), welches auf Erkenntnissen aus Forschungsarbeiten zur *Organizational Justice* basiert und somit auch die Aspekte der Nachvollziehbarkeit und der wissenschaftlichen Überprüfbarkeit erfüllt und welches ich im nächsten Kapitel ausführlich darstelle.

5.4 Das Modell der Bewerberreaktionen auf Personalauswahlverfahren von Gilliland

Gilliland veröffentlichte 1993 ein Modell zur wahrgenommenen Fairness in Personalauswahlverfahren, welches auf Theorien und Erkenntnissen zur organisationalen Gerechtigkeit aufbaut. Damit stellte er anderen Forschern ein theoriebasiertes Gerüst für weitere Forschungsarbeiten zur Verfügung, was dazu führte, dass die Organizational Justice Theory das am häufigste verwendete Paradigma zur Erforschung und Erklärung von Bewerberreaktionen wurde (Schleicher et al., 2006). Zudem ist es Gilliands Verdienst, dass zusätzlich zu den psychometrischen Aspekten eines Testverfahrens heute auch dessen Akzeptanz bei den Bewerbern als wichtiges Gütekriterium gilt und die damit verbundenen möglichen Auswirkungen auf die Organisation bekannt sind.

Bei der organisationalen Gerechtigkeit liegt der Fokus auf der Wahrnehmung von Fairness im Arbeitsumfeld und den entsprechenden Reaktionen darauf (Ployhart & Ryan, 1997). Der Begriff geht auf Greenberg (1987) zurück, der diesen als Sammelbezeichnung für verschiedene Fairnesskonzepte wählte. In der heute allgemein akzeptierten Form umfasst das Paradigma vier Gerechtigkeitsformen: Die Gerechtigkeit von Verteilungsregeln für Güter und der daraus resultierenden Ergebnisse (Verteilungsgerechtigkeit, *distributive justice*), die Gerechtigkeit des Vorgehens und der eingesetzten Verfahren bei der Verteilung von Gütern (Verfahrensgerechtigkeit, *procedural justice*), die Qualität der Behandlung der beteiligten Personen bei der Durchführung der Verfahren (interpersonale

Gerechtigkeit, *interpersonal justice*) und die Angemessenheit der Informationen zur Art der Durchführung der Verfahren und der Entscheidungsfindung (informationale Gerechtigkeit, *informational justice*) (z. B. Bell et al., 2006; Colquitt, 2001; Greenberg, 1993a; Nowakowski & Conlon, 2005). Dem Konzept der organisationalen Gerechtigkeit und seinen Gerechtigkeitsformen kommt im Bereich der Arbeitspsychologie eine grosse theoretische wie praktische Bedeutung zu: In metaanalytischen Studien wiesen Cohen-Charash und Spector (2001) und Colquitt et al. (2001) nach, dass organisationale Gerechtigkeit mit arbeitsrelevanten Einstellungen und Verhaltensweisen wie Arbeitszufriedenheit, Leistungsbereitschaft, Organizational Citizenship Behavior, Kündigungsabsichten oder Mitarbeiterkriminalität zusammenhängt.

Nachfolgend stelle ich die vier Gerechtigkeitsformen kurz vor. Für ausführlichere Übersichtsartikel verweise ich auf Colquitt et al. (2001), Cropanzano, Bowen und Gilliland (2007) oder Greenberg (1990). Einen Überblick über die historischen Wurzeln bieten Byrne und Cropanzano (2001) oder Colquitt, Greenberg und Zapata-Phelan (2005).

Verteilungsgerechtigkeit

Das Konzept der Verteilungsgerechtigkeit beschreibt im Arbeitskontext die Wahrnehmung der Fairness der Verteilung von Vergütungen und Anreizen wie beispielsweise Lohn, Lob, Beförderungen, Bonifikationen oder Aufträgen. Es handelt sich dabei um die Basis der Beziehung Arbeitgeber-Arbeitnehmer, in welcher Arbeit gegen Lohn ausgetauscht wird und sich der Beschäftigte somit zu Recht die Frage stellt, ob er das bekommt, was ihm auf Grund seines Einsatzes für die Firma auch zusteht und ob dies vergleichbar mit der Entlöhnung und Belohnung der anderen Beschäftigten ist. Das Konzept geht auf Arbeiten von Stouffer, Suchman, DeVinney, Star und Williams (1949, Ungerechtigkeitsempfinden bei Beförderungen im Militär; *relative deprivation concept*) und Homans (1961, Gerechtigkeit im sozialen Austausch) zurück (Colquitt et al., 2005). Adams (1963) nahm die Ideen von Homans auf und entwickelte diese weiter zu seiner Equity-Theorie (Gleichgewichtstheorie), welche besagt, dass eine Person das Verhältnis zwischen dem erbrachten Beitrag und der erhaltenen Belohnung als fair erachtet, wenn es gleich ist, wie das anderer Personen in vergleichbaren Situationen. Deutsch (1975) und Leventhal (1976) erweiterten die Theorie von Adams um die beiden Aspekte Gleichheit (*equality*: jedem gleich viel) und Bedürfnis (*need*: jedem das, was er am dringendsten benötigt).

Verfahrensgerechtigkeit

Leventhal (1980) kritisierte, dass bei der Equity-Theorie der Prozess, welcher zum Verteilungsergebnis und somit zum Fairnessempfinden führt, keine Berücksichtigung findet. Er erkannte, dass Personen auch diesen Prozess, unabhängig vom Ergebnis, als fair oder unfair erleben und bezeichnete ihn, Bezug nehmend auf die Arbeit von Thibaut und Walker (1975), als prozedurale Fairness. Letztere untersuchten die Reaktionen auf Kontrollmöglichkeiten über den Verlauf von Gerichtsprozessen und stellten in Experimenten fest, dass die Angeklagten dann einen Prozess als fair empfinden, wenn sie Argumente zur Verteidigung vorbringen und ihre Ansichten äussern können (*process control*) und sie den Eindruck haben, dass sie einen Einfluss auf den Ausgang des Prozesses haben (*decision control*). Leventhal (1980, siehe auch Leventhal, Karuza & Fry, 1980) postulierte sechs Regeln zur Überprüfung der Fairness eines Prozesses: Konsistenz des Verfahrens über Personen und Zeit; Unvoreingenommenheit der Beteiligten; Akkuratheit der zugrunde liegenden Informationen; Möglichkeit zur Korrektur von Entscheidungen; Widerspiegeln der Anliegen, Werte und Ansichten aller von der Entscheidung Betroffenen; Berücksichtigung der ethischen und moralischen Werte der Beteiligten. Eine Konsolidierung fand das Konzept schliesslich in der Monografie von Lind und Tyler (1988).

Interpersonale und informationale Gerechtigkeit (interaktionale Gerechtigkeit)

Bies und Moag (1986) beschrieben das Konzept der interaktionalen Gerechtigkeit, welche sich auf die wahrgenommene Gerechtigkeit der zwischenmenschlichen Behandlung durch eine Autorität bezieht. Aufgrund von Studien zu den Erwartungen der Bewerber zur Behandlung während eines Selektionsverfahrens identifizierten sie vier Kriterien interaktionaler Gerechtigkeit: Begründung, Aufrichtigkeit, Respekt und Korrektheit. Diese lassen sich in den zwei Dimensionen Erklärungen und Sensitivität zusammenfassen (z. B. Colquitt, 2001; Greenberg, 1990), für welche später Greenberg (1993a, 1993b) die Begriffe interpersonale und informationale Gerechtigkeit einführte. Interpersonale Gerechtigkeit umfasst den Grad, mit welchem Autoritätspersonen Arbeitnehmer mit Freundlichkeit, Würde und Respekt behandeln. Die Wahrnehmung der informationalen Gerechtigkeit beruht auf den Erklärungen zum Sinn und Zweck von Abläufen und Prozessen und zur Art und Weise derer Durchführung und zu den Verteilungsregeln der Vergütungen und Anreize. Anhand dieser Definitionen lässt sich auch der Unterschied zur Verfahrensgerechtigkeit ableiten: Bei letzterer bezieht sich das Fairnessurteil des Arbeitnehmers eher auf die Organisation als Ganzes, bei der

interaktionalen Gerechtigkeit auf das Verhalten einzelner Personen, zum Beispiel das des Vorgesetzten (Cohen-Charash & Spector, 2001).

Das Gerechtigkeitsmodell mit den vier Faktoren, zu welchem Colquitt (2001) ein Messinstrument entwickelt hat (siehe auch Colquitt & Shaw, 2005; Maier, Streicher, Jonas & Woschée, 2007), konnte mehrfach bestätigt werden (z. B. Colquitt, 2001; Colquitt et al., 2001; Judge & Colquitt, 2004; Maier et al., 2007; Streicher, Jonas, Maier, Frey, Woschée & Wassmer, 2008) und hat sich nun nach einer jahrelangen Debatte über die Eigenständigkeit der interaktionalen Gerechtigkeit von der Verfahrensgerechtigkeit als Modell der organisationalen Gerechtigkeit etabliert (Bies, 2005). Dies wird auch durch die Ergebnisse meta-analytischer Studien gestützt, in welchen die Autoren trotz der mittleren bis hohen Korrelationen zwischen den vier Gerechtigkeitsformen – interpersonale und informationale Gerechtigkeit korrelieren zum Beispiel mit r = .64 (Colquitt, 2001) – jeweils unterschiedliche Zusammenhänge zu verschiedenen Aspekten arbeitsrelevanten Verhaltens nachweisen konnten (Cohen-Charash & Spector, 2001 (Dreifaktormodell); Colquitt et al., 2001 (Vierfaktormodell); Vergleich der beiden Studien in Nowakowski & Conlon, 2005). So wirkt sich die wahrgenommene Verfahrensgerechtigkeit auf die Arbeitsleistung, die Arbeitszufriedenheit und die Zufriedenheit mit dem Management aus, wohingegen die Verteilungsgerechtigkeit einen Einfluss auf die Zufriedenheit mit der Entlöhnung, die Kündigungsabsicht und die Zufriedenheit mit dem direkten Vorgesetzten hat. Die beiden interaktionalen Gerechtigkeitsformen unterscheiden sich dahingehend, dass sich die informale Gerechtigkeit stärker auf die Zufriedenheit mit dem Management und die Kündigungsabsicht auswirkt.

Besonders nützlich erwies sich das Modell der organisationalen Gerechtigkeit für die Analyse der Abläufe in der Personalselektion (Gilliland & Hale, 2005; Truxillo et al., 2004). Nachdem Folger und Greenberg 1985 die Wichtigkeit der Verfahrensgerechtigkeit für Forschung und Praxis im Personalbereich von Organisationen aufzeigten und damit eine rege Forschungs- und Publikationstätigkeit auslösten, publizierte Singer 1990 den ersten Artikel, welcher sich ausschliesslich dem Zusammenhang zwischen organisationaler Gerechtigkeit und Personalselektion widmete. Der Beitrag von Smither et al. (1993) zu Bewerberreaktionen markiert den Übergang der Erforschung der Theorien zum sozialen Prozess der Personalselektion (Herriot, 1989; Schuler, 1993a) zur Gerechtigkeitstheorie: Ihr Erhebungsinstrument umfasst sowohl Dimensionen aus dem Konzept der sozialen Validität als auch zur Verteilungs- und Verfahrensgerechtigkeit. Gilliland (1993; siehe auch Gilliland & Hale, 2005) entwickelte schliesslich ein auf der Organizational Justice Theory basierendes Modell der Bewerberreaktionen auf Personalselektionsverfahren (siehe Abbildung 5.1) und formulierte zehn Regeln

(siehe Tabelle 5.1), deren praktische Umsetzung dazu führen soll, dass die Bewerber ein Selektionsverfahren als fair einstufen. In einer Übersichtstabelle zeigt Gilliland (1993) die Zusammenhänge zwischen diesen Regeln und einigen Theorien zu organisationaler Gerechtigkeit (Greenberg, 1986; Leventhal, 1980; Sheppard & Lewicki, 1987; Thibaut & Walker, 1975; Tyler & Bies, 1990) und weiteren Publikationen zu Bewerberreaktionen (Arvey & Sackett, 1993; Iles & Robertson, 1989; Schuler, 1993a) auf. Dabei fällt eine grosse Übereinstimmung zur Kriteriensammlung von Arvey und Sackett (1993) auf, in welcher sechs der zehn Regeln enthalten sind. Die Publikation des Modells von Gilliland führte zu einer grossen Forschungsaktivität, so dass in den vergangenen 15 Jahren mehr als 200 Artikel erschienen sind, in welchen die Autoren darauf Bezug nehmen (für Reviews siehe Ryan & Ployhart, 2000; Truxillo et. al, 2004). Nachfolgend stelle ich das Modell und empirische Belege zu den einzelnen Regeln dar, wobei ich die von Gilliland (1993) vorgenommene Gruppierung der zehn Regeln in die Bereiche formale Charakteristiken, Erklärungen und zwischenmenschlicher Umgang übernehme.

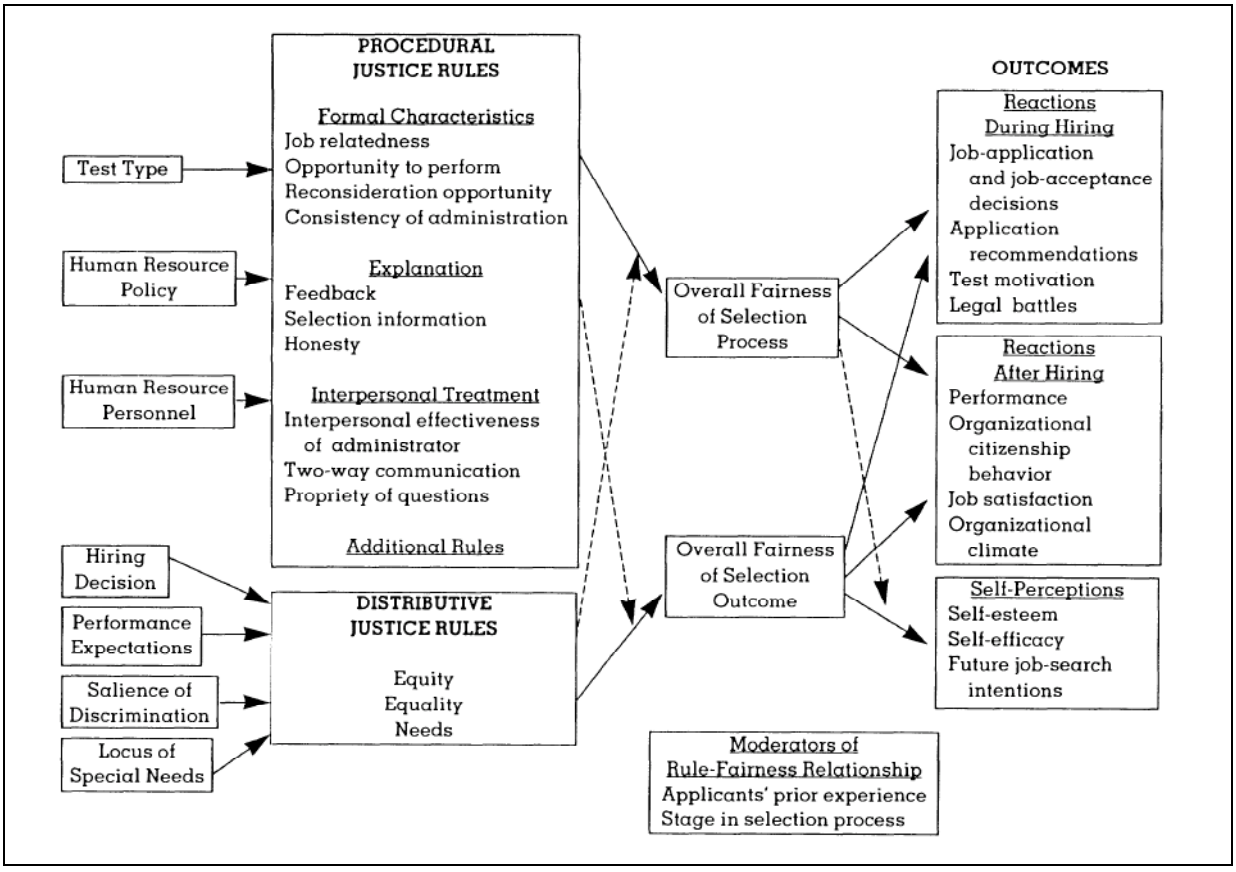


Abbildung 5.1 Modell der Bewerberreaktionen auf Personalselektionsverfahren (Gilliland, 1993, S. 700).

Tabelle 5.1

Regeln der Reaktionen auf Selektionsprozesse (nach Gilliland & Honig, 1994)

A) Formale Merkmale (formal characteristics)		
Job Relatedness	Tätigkeitsbezug	Ausmass, in welchem der Test tätigkeitsrelevante Fähigkeiten misst (Augenscheinvalidität).
Opportunity to Perform	Möglichkeit zur Selbstdarstellung	Adäquate Möglichkeit, sein Wissen, seine Fähigkeiten und seine Fertigkeiten unter Beweis zu stellen.
Reconsideration Opportunity	Möglichkeit zur Wiedererwägung	Möglichkeit, Testergebnisse anzuschauen, diese anzufechten oder eine Testwiederholung zu erwirken.
Consistency of Administration	Vergleichbarkeit der Durchführung	Standardisierung des Selektionsverfahrens.
B) Erklärung (explanation)		
Feedback	Ergebnisrückmeldung	Zeitgerechtes und aussagekräftiges Feedback zu den Testergebnissen und zum Selektionsentscheid.
Selection Information	Information zum Auswahlverfahren	Informationen und Begründungen zum Selektions- und Entscheidungsprozess.
Honesty	Aufrichtigkeit	Offene und aufrichtige Kommunikation mit dem Bewerber.
C) Zwischenmenschlicher Umgang (interpersonal treatment)		
Interpersonal Effectiveness	Respektvolle Behandlung	Warme und respektvolle Behandlung des Bewerbers.
Two–way Communication	Zweiweg-Kommunikation	Möglichkeit für den Bewerber, Fragen zu stellen und Anregungen zu geben.
Propriety of Questions	Angemessenheit der Fragen	Verzicht auf unangemessene, in die Persönlichkeitssphäre eindringende oder diskriminierende Fragen.

A) Formale Merkmale

Unter den formalen Aspekten der Gerechtigkeitswahrnehmung eines Selektionsprozesses subsumiert Gilliland (1993) in Anlehnung an Leventhal (1976, 1980) die vier Aspekte Tätigkeitsbezug, Möglichkeit zur Selbstdarstellung, Möglichkeit zur Wiedererwägung und Vergleichbarkeit der Durchführung.

Tätigkeitsbezug (job relatedness)

Hierunter versteht Gilliland das Ausmass, in welchem ein Verfahren Inhalte misst, die einen direkten und offensichtlichen Bezug zur Tätigkeit haben oder die valide zu sein scheinen, indem sie zukünftige Leistung am Arbeitsplatz vorhersagen können (also Inhalts- und prädiktive Validität). Dieser Aspekt ist als „Transparenz" schon im Modell von Schuler und Stehle (1983) und der Kriterienliste von Arvey und Sackett (1993) enthalten und hängt mit dem Konzept der Augenscheinvalidität zusammen (z. B. Mosier, 1947; Nevo, 1993).

Zu keiner anderen der zehn Regeln liegen so viele Forschungsergebnisse vor, wie zum Tätigkeitsbezug (Madigan & Macan, 2005; Schleicher et al., 2006) und es scheint sich um den wichtigsten Aspekt für das Empfinden von Fairness in einem Selektionsprozess zu handeln (Ryan & Ployhart, 2000; siehe auch Chan, Schmitt, Jennings, Clause & Delbridge, 1998; Gilliland, 1994; Madigan & Macan, 2005; Schmidt, Greenthal, Hunter, Berner & Seaton, 1977; Schmitt, Gilliland, Landis & Devine, 1993; Steiner & Gilliland, 1996; Truxillo, Bauer & Sanchez, 2001): Wenn der Bewerber den Tätigkeitsbezug nicht erkennen kann, wird er nicht nachvollziehen können, dass das Testverfahren relevante Arbeitsleistung vorhersagt und somit den Test als unfair erleben (Cropanzano & Wright, 2003). Dies ist auch empirisch bestätigt: Bewerber stufen Arbeitsproben, Interviews, Assessment Center- und Intelligenztests mit konkreten Inhalten als tätigkeitsbezogener und gerechter ein als Intelligenztests mit abstrakten Aufgaben, Persönlichkeits-Fragebogen oder graphologische oder astrologische Gutachten (z. B. Anderson & Witvliet, 2008; Chan et al., 1997; Gilliland, 1994; Kluger & Rothstein, 1993; Kravitz et al., 1996; Rynes & Connerley, 1993; Schmidt et al., 1977; Smither et al., 1993; Stone-Romero, Stone & Hyatt, 2003; siehe auch Kapitel 5.7). Dieser Befund gilt auch für Testverfahren derselben Kategorie: Ryan, Greguras und Ployhart (1996) zeigten auf, dass Bewerber bei sportlichen Leistungsprüfungen diejenigen Tests als fairer einstufen, welche einen höheren Bezug zur Tätigkeit haben. Hingegen scheinen unterschiedliche Darbietungsformen (zum Beispiel Video vs. Text) keinen Einfluss auf die Einschätzung des Tätigkeitsbezuges zu haben (Kanning, Grewe, Hollenberg & Hadouch, 2006). Die Wahrnehmung der prädiktiven Validität durch den Testbearbeiter kann sich auch auf die Leistung im Test auswirken: Chan et al. (1997) stellten fest, dass die Testleistung (IQ) umso höher ausfällt, je höher der Zusammenhang zwischen Test und Job wahrgenommen wird, was sie durch die mediierende Wirkung der Testmotivation erklären.

Ryan und Chan (1999) zeigten auf, dass es sich beim Tätigkeitsbezug (*job-related face validity*) und der prädiktiven Validität um zwei unterschiedliche Konstrukte handelt, welche aber beide hoch mit der Wahrnehmung der Verfahrensgerechtigkeit korrelieren. Bei der Konstruktion einer Skala zur Erfassung der zehn Gilliland-Faktoren durch Bauer et al. (2001) ergab die explorative Faktorenanalyse elf Faktoren, welche die Autoren auch konfirmatorisch überprüften, wobei sich *job relatedness/content* und *job relatedness/predictive* als zwei eigenständige Faktoren herausstellten – ein Befund, den 1993 schon Smither et al. publizierten. Auf Grund dieser Ergebnisse raten Gilliland und Hale (2005, S. 422) Forschern „to seperate perception of job-related face validity and predictive validity in their studies".

Möglichkeit zur Selbstdarstellung (opportunity to perform)

Bei der Möglichkeit zur Selbstdarstellung handelt es sich um die Wahrnehmung, dass einem als Bewerber während des Selektionsverfahrens eine angemessene Gelegenheit geboten wird, sein Wissen, seine Fähigkeiten und Fertigkeiten zu zeigen, um so wenigstens das Gefühl der Ausübung einer minimalen Kontrolle über den Selektionsprozess zu erleben (Schuler, 1993a; Thornton, 1993). Sie ist damit ein weiteres formales Merkmal des Tests oder des Verfahrens, das aber – im Gegensatz zum Tätigkeitsbezug – unabhängig von der zu besetzenden Stelle ist (Schleicher et al., 2006), da ein Bewerber durchaus den Eindruck haben kann, dass er die Gelegenheit zur Darstellung seiner Stärken hat, er jedoch den Zusammenhang zur zu besetzenden Stelle nicht sieht. Für Bewerber hat dieser Aspekt eine grosse Bedeutung für die Bewertung des Auswahlverfahrens, da sie ihn unmittelbar erleben. So beurteilen abgelehnte Bewerber die Möglichkeit zur Selbstdarstellung nach der Verkündung des Entscheides als den mit Abstand wichtigsten Faktor der Verfahrensgerechtigkeit noch vor dem Tätigkeitsbezug, der respektvollen Behandlung und der Zweiweg-Kommunikation (Schleicher et al., 2006), was durch einen selbstwertschützenden externalen Attributionsstil begründet ist (Arvey, Strickland, Drauden & Martin, 1990. Siehe auch Kelley & Michela, 1980; Mehlman & Snyder, 1985). Die praktische Relevanz dieser Regel liess sich zudem in verschiedenen Studien nachweisen: So zeigten zum Beispiel Truxillo et al. (2001) auf, dass Bewerber das Selektionsverfahren umso fairer einstufen und das Ergebnis besser akzeptieren, je mehr ihnen die Möglichkeit geboten wird, ihre Fähigkeiten zu demonstrieren (siehe auch Bies & Shapiro, 1988; Dipboye & de Pontbriand, 1981; Kluger & Rothstein, 1993). Dieser Aspekt liefert auch eine Erklärung dafür, dass Bewerber unstrukturierte Interviews gegenüber strukturierten bevorzugen (z. B. Conway & Peneno, 1999; Kohn & Dipboye, 1998; Latham & Finnegan, 1993; Schuler, 1993b).

Gemäss der Studie von Schleicher et al. (2006) sind die fünf am häufigsten genannten Merkmale der Möglichkeit zur Selbstdarstellung genügend Zeit (siehe dazu auch Singer, 1990), ausreichende Anweisungen und Hilfsmittel, Übereinstimmung zwischen dem Test respektive dem Verfahren und dem eigenen Wissen respektive der eigenen Erfahrung, die Angemessenheit des Testformates, um seine Fähigkeiten zu zeigen und die direkte Interaktion mit dem Assessor. Die Autoren stellen abschliessend fest, dass die Personalverantwortlichen diesen Aspekten in der Praxis zu wenig Aufmerksamkeit schenken und präsentieren eine Liste mit Vorschlägen, die dazu dienen, die Möglichkeiten zur Selbstdarstellung in Selektionsprozessen zu verbessern:

- Durchführen mindestens eines nichtschriftlichen Testverfahrens, wie zum Beispiel einer Übung.
- Einsatz eines Interviews, welches aus einem unstrukturierten und einem strukturierten Teil besteht.
- Anpassen des Inhalts des Interviews auf die Vorerfahrung des Bewerbers.
- Genügend Zeit und eine störungsfreie Umgebung anbieten.
- Ein auf den Background der Bewerberpopulation abgestimmtes Testmaterial einsetzen.
- Eingehen auf oben genannte Aspekte bei der Instruktion der Bewerber und dem Feedback.

Möglichkeit zur Wiedererwägung (reconsideration opportunity)

Gilliland (1993) beschreibt die Möglichkeit zur Wiedererwägung als die vom Bewerber wahrgenommene Beeinflussbarkeit des Entscheidungsprozesses, indem ihm der Personalverantwortliche seine Resultate zeigt und erläutert und ihm die Chance eröffnet, die Resultate zu überprüfen oder die Testung nochmals durchzuführen. Dieser Aspekt geht auf Leventhal (1976, 1980) zurück, welcher die Korrigierbarkeit eines Ergebnisses oder Entscheides als ein Merkmal eines fairen Prozesses aufführte und wurde von Arvey und Sackett (1993) als wichtiger Faktor für die Fairnessbeurteilung von Selektionsprozessen angesehen. Greenberg (1986) konnte in seiner Studie nachweisen, dass die Möglichkeit, Ergebnisse von Leistungsbeurteilungen anzufechten, eine Determinante eines als fair empfundenen Vorgehens ist. Gilliland (1993) regte jedoch an, diesen Bereich noch besser zu untersuchen und stellt zehn Jahre später fest, dass es unklar ist, ob eine zweite Chance mit einem vergleichbaren Test zu höherer Fairnesseinschätzung führt (Gilliland & Hale, 2005; siehe auch Gilliland, 1995). So fanden auch Dineen, Noe und Wang (2004) in ihrer Untersuchung, dass die Wiedererwägung bei der Fairnessbeurteilung eines webgestützten Screeningverfahrens nur einen geringen Stellenwert hat. Hingegen zeigte Graczyk (2005) auf, dass zwischen der Möglichkeit zur Wiedererwägung und der Attraktivität der Organisation, der Annahme eines Arbeitsangebotes, der Selbstwirksamkeitserwartung und der Wahrscheinlichkeit, Anklage zu erheben Zusammenhänge bestehen. Cropanzano und Wright (2003) nehmen an, dass die Möglichkeit zur Wiedererwägung für externe Bewerber auf eine Stelle nicht so wichtig ist, wohingegen diese für bereits angestellte Mitarbeiter zentral ist (siehe auch McEnrue, 1989).

Vergleichbarkeit der Durchführung (consistency of administration)

Die Regel der Vergleichbarkeit der Durchführung bezieht sich auf das Gefühl der Bewerber, dass der Selektionsprozess für alle gleich ist und grundsätzlich alle dieselben Chancen haben, die Stelle zu erhalten. Auch diese Regel beschrieben schon Arvey und Sackett (1993), welche als zusätzliches Merkmal die standardisierte Auswertung und Interpretation der Testscores aufführten, wobei sich hier die Frage stellt, auf Grund welcher Hinweise die Bewerber dies beurteilen können. Es ist deshalb anzunehmen, dass sich die Einschätzung der Gleichbehandlung hauptsächlich auf den Strukturierungsgrad der eingesetzten Verfahren und die Art der Tests abstützt. Weiter beurteilen Bewerber Organisationen als gerechter, wenn diese ein transparentes Selektionsverfahren einsetzen, das erkennen lässt, dass alle Bewerber gleich behandelt werden (Gilliland & Hale, 2005).

Die empirische Evidenz dieser Regel ist nicht ganz eindeutig: Neben dem Tätigkeitsbezug war in der Untersuchung von Ryan et al. (1996) noch die Vergleichbarkeit der Durchführung ein signifikanter, wenn auch schwacher Prädiktor für Fairness, bei Madigan und Macan (2005) war es von sechs Merkmalen der Verfahrensgerechtigkeit der zweitschwächste Einflussfaktor auf die Gesamteinschätzung der Fairness und bei Dineen et al. (2004) der stärkste von insgesamt fünf untersuchten Verfahrensmerkmalen. Ployhart und Ryan (1998) zeigten auf, dass Ungleichbehandlungen in der zur Verfügung stehenden Zeit bei einer Testdurchführung zu einer tieferen Einschätzung der Fairness bei den Benachteiligten führen. In der Studie von Ployhart, Holcombe Ehrhart und Hayes (2005) liess sich jedoch nur ein Effekt der Vergleichbarkeit der Durchführung (Zulassungsverfahren wird jedes Jahr geändert vs. Zulassungsverfahren bleibt immer dasselbe) in Kovariation mit der Variable Vergleichbarkeit mit anderen Zulassungsverfahren (Zulassungsverfahren ist gleich wie in anderen Universitäten) auf die wahrgenommene Prozessfairness nachweisen.

B) Erklärung

Unter dem Oberbegriff Erklärung subsumiert Gilliland (1993) die Informationen, welche der Personalverantwortliche dem Bewerber zum Selektionsprozess, zu seiner Leistung und seinen Ergebnissen abgibt. Es versteht sich von selbst, dass der Bewerber dabei den Eindruck erhalten muss, dass er offen und ehrlich informiert wird.

Ergebnisrückmeldung (feedback)

Tyler und Bies (1990) sehen in einem rechtzeitigen und informativen Feedback einen wichtigen Faktor der Wahrnehmung der interaktionalen Gerechtigkeit. Eine detaillierte Rückmeldung der Ergebnisse ist in einer Selektionssituation von grosser Bedeutung, wenn man im Sinne der bewerberzentrierten Personalauswahl (Boss, 2005) davon ausgeht, dass der Bewerber unabhängig vom Entscheid von der Teilnahme am Selektionsverfahren profitieren und so quasi eine Entschädigung für sein Engagement und die aufgewendete Zeit erhalten soll (Cropanzano & Wright, 2003). In diesem Sinne gibt der Personalverantwortliche idealerweise neben einem umfassenden und bewerberorientierten Feedback auch noch Entwicklungshinweise. Davon profitiert schlussendlich nicht nur der Bewerber, sondern indirekt auch das Unternehmen, da dies zu einer deutlich besseren Akzeptanz der Testverfahren und des Selektionsprozesses führt, wie Resultate aus mehreren Studien belegen (z. B. Gilliland, 1995; Gilliland et al., 2001; Horvath et al., 2000; Lounsbury, Bobrow & Jensen, 1989). Zudem wirkt sich ein von den Bewerbern gut akzeptiertes Feedback – in Übereinstimmung mit der *self-consistency theory* von Korman (1970) – auf die spätere Arbeitsleistung aus (Anseel & Lievens, 2009).

Die grosse Bedeutung des Feedbacks für die Fairnesswahrnehmung des gesamten Selektionsprozesses belegen auch Van Vianen, Taris, Scholten und Schinkel (2004) in ihrer Studie mit über 400 Bewerbern aus verschiedenen Organisationen: Das Erleben der Art und Weise des Feedbackgebens und der Vollständigkeit des Feedbacks beeinflussen die Fairnesswahrnehmung (*post feedback fairness perception*) eines aus drei Tests bestehenden Selektionsverfahrens bedeutend stärker, als der grundsätzliche Glaube an Testverfahren, der wahrgenommene Tätigkeitsbezug der Tests und die wahrgenommene Leistung im Selektionsverfahren (*prefeedback fairness perception*). Die von den Bewerbern beurteilte Attraktivität der Arbeitsstelle wurde zudem nur durch diese beiden Feedbackeinstufungen beeinflusst, nicht aber durch die Fairnesswahrnehmung.

Vorsicht ist jedoch bei abgewiesenen Bewerbern geboten, welche das Feedback als weniger zutreffend einstufen und somit schlechter akzeptieren, als aufgenommene (Anseel & Lievens, 2009; siehe auch Brett & Atwater, 2001): Schinkel et al. (2004) zeigen auf, dass sich in diesem Fall eine detaillierte Leistungsrückmeldung im Vergleich zu einer einfachen Absage negativ auf die Selbsteinschätzung und die emotionale Befindlichkeit auswirkt (siehe auch Ilgen & Davis, 2000; Ployhart et al., 2005; Ployhart et al., 1999). Wie dem entgegengewirkt werden kann, führen Dipboye und de Pontbriand (1981) am Beispiel von

Leistungsbeurteilungen aus: Mitarbeiter akzeptieren negatives Feedback besser, wenn sie aktiv am Feedbackprozess beteiligt sind, das weitere Vorgehen und die Zielvereinbarungen diskutieren können und sich die Beurteilung auf arbeitsrelevante Inhalte bezieht. Die Meta-Analyse von Shaw, Wild und Colquitt (2003) ergab zudem, dass Erklärungen in Form einer Entschuldigung besser akzeptiert werden, als wenn sie als Rechtfertigung vorgebracht werden.

Wichtig für die Fairnesseinstufung ist neben der Vollständigkeit des Feedbacks auch dessen Rechtzeitigkeit: Bei der Studie von Dineen et al. (2004) erwies sich der Faktor „rechtzeitiges Feedback" zwar als schwächster von fünf Prädiktoren für wahrgenommene Fairness. In einer anderen, qualitativen Studie berichteten die Bewerber jedoch, dass Verzögerungen im Rekrutierungsprozess der häufigste Grund für einen Rückzug der Bewerbung war, da diese ein Bild einer desorganisierten Organisation abgeben (Rynes, Bretz & Gerhart, 1991). Truxillo et al. (2002) konnten aufzeigen, dass eine Erklärung für ein verzögertes Feedback negative Auswirkungen auf die Fairnesseinschätzung abschwächt.

Truxillo et al. (2002) schlagen vor, dass auf Grund des Einflusses eines ausführlichen und zutreffenden Feedbacks auf die Wahrnehmung der interpersonalen Fairness, die Wahrnehmung der Attraktivität der Organisation (Anseel & Lievens, 2009) oder das Organizational Citizenship Behavior Personalverantwortliche im Feedbackgeben zu schulen seien.

Information zum Auswahlverfahren (selection information)

Informationen und Erklärungen zum Selektionsprozess, zu den eingesetzten Testverfahren und zur Entscheidungsfindung tragen dazu bei, das Gefühl der Unsicherheit der Bewerber zu senken und führen dazu, dass sich die Bewerber fairer behandelt fühlen (Shaw et al., 2003) und die Testverfahren besser akzeptieren (z. B. Rolland & Steiner, 2007). So zeigten Borchers (1986) und Lülsdorf (1986, beide zitiert nach Schuler, 1990) in ihren empirischen Studien auf, dass sich detaillierte Informationen über das eingesetzte Selektionsverfahren positiv auf das Akzeptanzurteil der Bewerber auswirken. Mit dem Geben umfassender Informationen soll erreicht werden, dass der Bewerber den Prozess versteht, weiss, auf Grund welcher Informationen der Personalverantwortliche oder der Linienvorgesetzte den Einstellungsentscheid trifft, weiss, weshalb welche Verfahren eingesetzt werden und dass keine moralisch-ethischen Standards verletzt werden (Cropanzano & Wright, 2003). Arvey und Sackett (1993) weisen darauf hin, dass die Reduktion von Ungewissheit durch das Geben von Informationen schliesslich auch dazu führt, dass der Bewerber ein schlechtes Abschneiden eher auf sich selbst bezieht, als auf die Testsituation oder die Organisation.

Als besonders wichtig für eine positive Fairnesswahrnehmung hat sich die Erklärung des Selektionsentscheides herausgestellt (Bies & Shapiro, 1987, 1988; Bies, Shapiro & Cummings, 1988; Colquitt & Chertkoff, 2002; Daly & Geyer, 1994; Ployhart et al., 1999), wobei dieser Effekt in einzelnen Studien nicht nachgewiesen werden konnte (Gilliland, 1994; Schaubroeck, May & Brown, 1994). Eine für den Bewerber nachvollziehbare Begründung ist jedoch insbesondere dann wichtig, wenn der Personalverantwortliche einen negativen Entscheid kommunizieren muss (Bies & Moag, 1986; Gilliland et al., 2001). Ployhart et al. (2005) zeigten zudem auf, dass aufgenommene und abgewiesene Bewerber unterschiedliche Informationen verwenden, um eine positive Selbstwahrnehmung aufrecht zu erhalten und die Fairness des Selektionsprozesses einzuschätzen: Wenn die Bewerber zum Beispiel wissen, dass die Selektionsquote tief ist, dass also nur wenigen eine Stelle angeboten wird, stuft diese Information ein abgewiesener Bewerber als Zeichen eines unfairen Prozesses ein, ein gewählter jedoch als ein solches eines fairen. Auch bei Burns et al. (2008) zeigten sich Unterschiede zwischen aufgenommenen und abgewiesenen Bewerbern: Sie gaben in einer Realsituation einem Teil der Bewerber Informationen zum Anforderungsprofil, zum Ablauf der Testung, zu den einzelnen Testverfahren, zum Zusammenhang der Tests zum Anforderungsprofil, zu Testbearbeitungs-Strategien und führten Itembeispiele und Vorbereitungshinweise auf. Die abgewiesenen Bewerber, welche die Testinformationen erhalten haben, waren deutlich zufriedener mit dem Selektionsverfahren und beurteilten dieses als fairer, als abgewiesene Bewerber, welche keine Testinformationen erhalten haben. Bei den aufgenommenen Bewerbern waren diese Zusammenhänge umgekehrt, ohne jedoch signifikant zu sein. Auf die Leistungen in den Testverfahren und die wahrgenommene Validität der Verfahren hatte diese Vorinformation keinen Einfluss. Dieses Ergebnis steht im Gegensatz zum Ergebnis von Clause, Delbridge, Schmitt, Chan und Jennings (2001), ist aber konsistent mit demjenigen von Ryan, Ployhart und Greguras (1998).

Truxillo et al. (2002) und Noon (2006) wiesen nach, dass ausführliche Informationen – nur der Hinweis auf die Reliabilität und die Validität reicht nicht aus (Lievens, De Corte & Brysse, 2003) – zu einem Test dazu führen, dass Bewerber diesen, das Auswahlverfahren insgesamt und auch die Organisation fairer einstufen. Überlagert wird dieser Effekt jedoch durch den grundsätzlichen Glauben an Testverfahren, welcher ein substanzieller indirekter Effekt auf die Fairnessbeurteilung aufweist (Lievens et al., 2003). Es scheint sich dabei um eine präventive, selbstwertdienliche Beurteilung (*self-serving bias*) zu handeln, welche gegen die Selbstwertbedrohung eines enttäuschenden Testergebnisses wirkt, was vor allem dann auftritt, wenn der Bewerber noch keine Erfahrungen mit

Testverfahren sammeln konnte und so keine Anhaltspunkte dafür hat, wie er im Test abschneiden wird (Van Vianen et al., 2004; siehe auch Bauer et al., 1998). In ihrer Meta-Analyse konnten Truxillo, Bodner, Bertolino, Bauer und Yonce (2009) zudem aufzeigen, dass Informationen zur Testdurchführung und zum Testinhalt auch die Leistung in einem Intelligenztest beeinflussen, wobei die Testmotivation diesen Zusammenhang mediiert.

Wie komplex die Zusammenhänge zwischen dem Geben von Informationen und Erklärungen und der Fairnesswahrnehmung jedoch sind, zeigt die Studie von Ambrose und Rosse (2003): Wenn sie Studenten erklärten, dass der Einsatz eines Persönlichkeitstests für die Besetzung von Praktikumsstellen typisch ist und sie gleichzeitig noch das Bedauern ausdrückten, dass einige Fragen sehr persönlich sind, jedoch keine Erklärung dafür lieferten wieso sie das Verfahren einsetzen, fiel die Einstufung der Fairness des Selektionsprozesses deutlich geringer aus, als wenn sie kein Bedauern ausdrückten (M = 4.85 vs. M = 3.63 auf einer Skala von 1 bis 7). In der Studie von LaHuis, Perreault und Ferguson (2003) zeigte sich, dass kurz gehaltene, allgemeine Hinweise zu einem Intelligenztest – nicht jedoch bei einem Persönlichkeits-Fragebogen – dessen Fairnesseinstufung erhöhten, wohingegen dies bei detaillierten Informationen nicht der Fall war. Schliesslich soll auch die Persönlichkeit – insbesondere die Selbstwirksamkeitserwartung – einen Einfluss auf die Fairnessbeurteilung haben (Horvath et al., 2000).

Aufrichtigkeit (honesty)

Die Regel Aufrichtigkeit behandelt das Ausmass, in welchem die Bewerber die Kommunikationsinhalte als ehrlich, offen, seriös und glaubwürdig einstufen. Unaufrichtigkeit bei der Darstellung der Organisation oder des zu besetzenden Arbeitsplatzes sind gemäss Bies und Moag (1986) die von den Bewerbern am häufigsten genannte Ursache für die Wahrnehmung von Unfairness und können dazu führen, dass ein Bewerber seine Bewerbung zurückzieht (Schmitt & Coyle, 1976).

Schon 1956 veröffentlichte Weitz die erste Studie zu den Auswirkungen des Gebens von realistischen Tätigkeitsinformationen (*realistic job preview*, RJP) auf die Einstellung der Bewerber gegenüber der Organisation und seither gehört dieses Thema zu den am meisten untersuchten Themen des Anstellungsprozesses (Rynes, 1991). Bei der RJP gibt der Personalverantwortliche dem Bewerber eine detaillierte Auflistung wichtiger, positiver sowie negativer Aspekte der zu besetzenden Stelle, was vor allem die Selbstselektion fördert (Schuler, 1990; Taylor & Bergmann, 1987; Wanous, 1992) und somit dazu beiträgt, dass die

Kündigungsrate bei neu eingestellten Mitarbeitern sinkt (McEvoy & Cascio, 1985; Meglino, Ravlin & DeNisi, 2000; Phillips, 1998). Jedoch fanden Bretz und Judge (1998) in ihrer Studie Hinweise darauf, dass zu Beginn des Selektionsprozesses gegebene negative Hinweise zur zu besetzenden Stelle vor allem bei sehr gut qualifizierten Bewerbern das Interesse an der fraglichen Stelle schwinden lassen. Eine unkluge Strategie wäre es nun aber, dem Bewerbern ausschliesslich Positives über die Arbeitsstelle und die Organisation zu berichten: Thorsteinson, Palmer, Wulff und Anderson (2004) konnten in einem Experiment zeigen, dass von zwei Firmendarstellungen die realistische die Organisation bei den Bewerbern in einem positiveren Licht erschienen liess, als die ausschliesslich positiv gehaltene. Dies führt zu folgendem Effekt: Realistische Darstellungen des Arbeitsplatzes führen zwar dazu, dass einige Bewerber ihr Interesse an der Stelle verlieren, dafür steigt die Chance, dass der schlussendlich aus den Reihen der Verbleibenden eingestellte Bewerber der Organisation länger treu bleibt. Meglino et al. (2000) wiesen nach, dass sich RJP nur dann schädigend auf das Unternehmen auswirken, wenn der Bewerber die zu besetzende Arbeitsstelle aus eigener Erfahrung kennt. Trifft dies nicht zu, nehmen die Bewerber ein Jobangebot eher an, wenn sie eine RJP bekamen, als wenn nur das Notwendigste an Informationen abgegeben wurde.

Premack und Wanous (1985) verwendeten in ihrer Meta-Analyse von 21 Studien acht Kriterien zur Erfassung der Auswirkungen von RJPs auf neu eingestellte Mitarbeiter: Wahrgenommenes Klima in der Organisation, Commitment gegenüber der Organisation, Einschätzung der eigenen Copingmöglichkeiten, Erwartungen, Arbeitszufriedenheit, Arbeitsleistung, Selbst-Selektion während des Bewerbungsprozesses und Verbleiben in der Organisation. Die Analyse zeigte, dass der Einfluss der RJP auf die Wahrnehmungen, Einstellungen und das Verhalten im Job gering ist. Die am stärksten beeinflussten Kriterien sind jedoch auch die wichtigsten: Das Verbleiben in der Organisation und die Arbeitsleistung.

C) Zwischenmenschlicher Umgang

Wie wichtig ein freundlicher, warmherziger und respektvoller Umgang mit den Bewerbern ist, belegt Rynes (1993), indem sie anekdotisch von Begegnungen mit Personalverantwortlichen berichtet, welche so unangenehm waren, dass die Bewerber daraufhin ihre Bewerbung zurückgezogen haben. Da diese die Personalverantwortlichen als stellvertretend für die Organisation wahrnehmen, ist es nahe liegend, dass dieser Kontakt auch einen Eindruck über die Organisation vermittelt (Bies & Moag, 1986; Iles & Robertson, 1989; Rynes, 1993).

Respektvolle Behandlung (interpersonal effectiveness)

Eine respektvolle und höfliche Behandlung der Bewerber führt dazu, dass sich diese wohl fühlen und im Auswahlverfahren eine optimale Leistung zeigen können. Dies gilt für die gesamte Untersuchungssituation, also auch für eventuell eingesetzte Testassistenten oder die Gestaltung der Untersuchungsräume. Rynes und Miller (1983) zeigten auf, dass eine positive Grundhaltung (Augenkontakt, Lächeln, ermutigendes Kopfnicken) dazu führt, dass die Bewerber denken, dass sie den Job erhalten, dass die Personalfachkraft die Unternehmung gut repräsentiert und dass in der Unternehmung die Mitarbeiter gut behandelt werden. In der Studie von Harn und Thornton (1985) zeigte sich, dass die Bewerber den Selektionsprozess besser akzeptieren, wenn der Personalverantwortliche sich für deren Gefühle interessiert, aktiv auf deren Aussagen eingeht und eine engagierte Diskussion mit ihnen führt. Verhält er sich im Interview jedoch wenig rücksichtsvoll und einfühlsam, kann dies bei den Bewerbern zu einer schlechten Beurteilung der Organisation führen (Liden & Parsons, 1986; Schmitt & Coyle, 1976).

Zweiweg-Kommunikation (two-way communication)

Gemäss den Studien von Singer (1990) und Madigan und Macan (2005) stellt für die Bewerber der gegenseitige Austausch von Informationen einen wichtigen Aspekt für das Empfinden der Fairness des gesamten Selektionsverfahrens dar. Es zahlt sich aus Sicht der Organisation demnach aus, wenn der Personalverantwortliche dem Bewerber die Möglichkeit bietet, Fragen zum Selektionsprozess, zur Arbeitsstelle und zur Unternehmung zu stellen. Zudem fördert eine offene Gesprächssituation die Äusserung der Bedenken und Anregungen zum Auswahlprozess. Diese sind dann auch angemessen zu berücksichtigen. Eine solche Haltung seitens des Personalverantwortlichen wirkt dem Gefühl des Ausgeliefertseins entgegen.

Dieser Aspekt liefert zudem eine Erklärung dafür, weshalb Bewerber unstrukturierte Interviews trotz geringerer Validität besser akzeptieren als strukturierte: Da erstere mehr Gestaltungsmöglichkeiten bieten, haben sie den Eindruck, sie könnten das Ergebnis eher beeinflussen (Kohn & Dipboye, 1998; Latham & Finnegan, 1993). Somit sollte bei strukturierten Interviews immer auch ein unstrukturierter Teil eingebaut werden, um die Fairnesswahrnehmung zu erhöhen (Gilliland & Hale, 2005). Zudem bevorzugen die Bewerber Interviewer, welche über gute Zuhörfähigkeiten verfügen (Harn & Thornton, 1985).

Angemessenheit der Fragen (propriety of questions)

Im Obligationenrecht Artikel 328b ist festgelegt, dass der Arbeitgeber nur Daten über den Arbeitnehmer bearbeiten darf, soweit sie dessen Eignung für das Arbeitsverhältnis betreffen. Somit ist auf gesetzlicher Ebene geregelt, dass die Privatsphäre des Bewerbers zu respektieren ist. Dem Bewerber steht sogar die gesetzlich erlaubte Notlüge zu, wenn er sich mit unzulässigen Fragen konfrontiert sieht (sog. „Notwehrrecht der Lüge"). Stellt ein Personalverantwortlicher trotzdem Fragen, welche eindeutig in die Privatsphäre des Bewerbers eindringen, zeigen verschiedenste Studien, dass dies einen äusserst negativen Effekt auf die Einstellung des Bewerbers zum Selektionsprozess und zur Unternehmung hat (Bies, 1993; Kravitz et al., 1996; Steiner & Gilliland, 1996; Stone & Stone, 1990).

Die zum Teil zu persönlichen Fragen in Persönlichkeits-Fragebogen, biografischen Fragebogen und Integritätstests sind mit ein Grund, weshalb Bewerber diese Verfahren schlechter akzeptieren, als zum Beispiel ein Interview oder einen Intelligenztest (Jones, 1991; Rosse et al., 1996; Smither et al., 1993; siehe auch Kapitel 5.7). Ganz schlecht bewerten Bewerber Verfahren, bei welchen sie keine Kontrolle mehr darüber haben, welche Informationen damit über ihre Persönlichkeit erfasst werden, wie dies auf die Graphologie oder die Lügendetektoren zutrifft (z. B. Kravitz et al., 1996).

Bauer et al. (2001) haben die zehn Gilliland-Dimensionen in einem Fragebogen operationalisiert und zwei übergeordnete Faktoren gefunden – soziale und strukturelle Fairness –, was kongruent zu den Aussagen von Greenberg (1990) ist. Weiterführende Studien zeigten auf, dass diese beiden Faktoren mit wichtigen organisationalen Einflussgrössen zusammenhängen (Bauer, Truxillo & Paronto, 2003). Bauer et al. (1998) haben zudem untersucht, welche der zehn Regeln in einer Testsituation von Bedeutung sind und konnten folgende identifizieren: Informationen über den Test; Möglichkeit, relevante Fähigkeiten während der Testung zu zeigen; Erhalten einer guten Behandlung während der Testdurchführung, mit der Möglichkeit, Fragen zu stellen; Konsistente Betreuung während der Testung; Einsatz von Testverfahren, welche einen Tätigkeitsbezug aufweisen.

Gilliland und Hale (2005) erweiterten die Liste mit den Regeln von Gilliland und Honig (1994) und ordneten die einzelnen Aspekte der organisationalen Gerechtigkeit den drei Phasen des Selektionsprozesses zu. Folgende Aspekte sind neu hinzugekommen: Leichtigkeit, Antworten zu verfälschen, verbreiteter Einsatz und Rechtzeitigkeit der Rückmeldung. Den Aspekt der Angemessenheit der Fragen teilen die Autoren neu in „Anständigkeit der Fragen" und „Eindringen in

die Privatsphäre" auf und subsumieren diese Aspekte unter der Verfahrensgerechtigkeit. Zudem führen sie neu die Verteilungsgerechtigkeitsaspekte *equity*, *equality* und *need* auf. In Tabelle 5.2 stelle ich die ursprüngliche Liste von 1994 der modifizierten von 2005 gegenüber.

Tabelle 5.2

Gegenüberstellung der Listen mit den Regeln der Reaktionen auf Selektionsprozesse von Gilliland und Honig (1994) und Gilliland und Hale (2005)

Gilliland und Honig (1994)	Gilliland und Hale, 2005
A) Formale Merkmale	*Verfahrensgerechtigkeit*
Tätigkeitsbezug	Tätigkeitsbezug (*job relatedness*)
Möglichkeit zur Selbstdarstellung	Möglichkeit zur Selbstdarstellung (*opportunity to perform*)
Möglichkeit zur Wiedererwägung	Möglichkeit zur Wiedererwägung (*reconsideration opport.*)
Vergleichbarkeit der Durchführung	Vergleichbarkeit der Durchführung (*consistency*)
	Verfälschbarkeit (*ease of faking answers*)
	Verbreiteter Einsatz (*widespread use*)
	Anständigkeit der Fragen (*propriety*)
	Eindringen in die Privatsphäre (*invasiveness*)
B) Erklärung	*Informationale Gerechtigkeit*
Ergebnisrückmeldung	Rechtzeitigkeit der Rückmeldung (*timeliness*)
Information zum Auswahlverfahren	Information zum Auswahlverfahren (*selection information*)
Aufrichtigkeit	Aufrichtigkeit (*honesty*)
C) Zwischenmenschlicher Umgang	*Interpersonale Gerechtigkeit*
Respektvolle Behandlung	Zwischenmenschlicher Umgang (*interpersonal effectiveness*)
Zweiweg-Kommunikation	Zweiweg-Kommunikation (*two-way communication*)
Angemessenheit der Fragen	
	Verteilungsgerechtigkeit
	Vergleichbarkeit (*equity*)
	Gleichheit (*equality*)
	Bedürfnis (*need*)

Die nachfolgende Tabelle 5.3 stellt die Zuordnung der Gerechtigkeitsaspekte von Gilliland und Hale zu den drei Stufen im Selektionsprozess dar: Recruiting und erste Kontaktaufnahme, Screening und Durchführen der Selektionsverfahren und Entscheidungsfindung und Entscheidungskommunikation. Aspekte der interpersonalen und informationalen Gerechtigkeit spielen dabei in allen drei Phasen eine Rolle, Verfahrensgerechtigkeit hauptsächlich beim Screening und beim Einsatz der Selektionsverfahren und die Verteilungsgerechtigkeit während der letzten Phase.

Tabelle 5.3

Gerechtigkeitsprinzipien bei Selektionsverfahren (nach Gilliland & Hale, 2005, S. 417)

Stufe im Selektionsprozess	Gerechtigkeits-Typ			
	Interpersonale Gerechtigkeit	Informationale Gerechtigkeit	Verfahrensgerechtigkeit	Verteilungsgerechtigkeit
Rekrutierung und erste Kontaktaufnahme	zwischenmenschlicher Umgang	Rechtzeitigkeit der Rückmeldung Information zum Auswahlverfahren Aufrichtigkeit	verbreiteter Einsatz Anständigkeit der Fragen	[nicht anwendbar]
Screening und Durchführung der Selektionsverfahren	zwischenmenschlicher Umgang Zweiweg-Kommunikation	Rechtzeitigkeit der Rückmeldung Information zum Auswahlverfahren Aufrichtigkeit	Tätigkeitsbezug Vergleichbarkeit der Durchführung Anständigkeit der Fragen Eindringen in die Privatsphäre Möglichkeit zur Selbstdarstellung Verfälschbarkeit Möglichkeit zur Wiedererwägung verbreiteter Einsatz	Vergleichbarkeit Bedürfnis
Entscheidungsfindung und Kommunikation der Entscheidung	zwischenmenschlicher Umgang	Rechtzeitigkeit der Rückmeldung Information zum Auswahlverfahren Aufrichtigkeit	[nicht anwendbar]	Vergleichbarkeit Gleichheit

Wie schon erwähnt, hat Gilliland mit seinem Modell der Bewerberreaktionen eine grosse Forschungsaktivität ausgelöst, welche auch neue Modelle hervorbrachte. Zwei davon stelle ich im nächsten Kapitel vor.

5.5 Weitere Modelle der Bewerberreaktionen

Nachfolgend stelle ich noch zwei weitere Modelle zu den Reaktionen von Bewerbern auf den Selektionsprozess dar: Das integrative Modell der Bewerberreaktionen auf Personalauswahlverfahren von Hausknecht, Day und Thomas (2004) und eine Darstellung der Struktur der Bewerberreaktionen aus dem militärischen Umfeld von Derous und Schreurs (2009).

5.5.1 Das integrative Modell der Bewerberreaktionen auf Personalauswahlverfahren von Hausknecht, Day und Thomas

Auf der Grundlage der Modelle von Gilliland (1993), Ryan und Ployhart (2000) und eines umfangreichen Literaturstudiums haben Hausknecht et al., (2004) ein Modell erarbeitet, in welchem sie die verschiedenen bisher in Forschungsarbeiten referierten Denkansätze und Modelle zu den Reaktionen von Bewerbern auf Selektionsverfahren integrieren. Dabei folgten sie der Aufforderung von Ryan und Ployhart, indem sie zur Erklärung der Wahrnehmung der Selektionssituation durch die Bewerber zusätzliche Erklärungsfaktoren hinzuzogen, welche über die Gerechtigkeitstheorie hinausgehen. In ihrem Modell unterscheiden sie – basierend auf Ryan und Ployhart – zwischen vier Kategorien von Variablen: Den Ausgangsbedingungen (Eigenschaften der Person, wahrgenommene Prozessmerkmale, Eigenschaften des Jobs und organisationaler Kontext), den Wahrnehmungen des Bewerbers während des Selektionsprozesses, den Auswirkungen der Wahrnehmungen des Bewerbers (Leistung im Selektionsverfahren, Selbstwahrnehmungen, Arbeitseinstellungen und Einstellungen und Verhalten gegenüber der Organisation) und Moderator-Variablen (z. B. Erwartungen, Selektionskontext, Alternativen), welche den Zusammenhang zwischen den Ausgangsbedingungen und den Wahrnehmungen des Bewerbers einerseits und den wahrgenommenen Prozessmerkmalen und den Auswirkungen andererseits beeinflussen. Das Modell habe ich in Abbildung 5.2 dargestellt.

Die Zusammenhänge zwischen den einzelnen Variablen eines solchen Modells sind sehr komplex. So fanden zum Beispiel Robertson, Iles, Gratton und Sharpley (1991) in ihrer Untersuchung, dass zwei Assessment Center, welche in einer Firma für verschiedene Teilnehmerkreise eingesetzt werden, nicht zu einheitlichen Reaktionen bei aufgenommenen und abgewiesenen Teilnehmern führen: Nur eines davon scheint so ausgestaltet zu sein, dass es beide Gruppen gleich gut akzeptieren. Zudem stellte sich heraus, dass Personen in unterschied-

lichen Karrierestufen verschieden auf Testverfahren reagieren, indem bei jüngeren Teilnehmern die vermutete Auswirkung auf den Karriereverlauf die Kündigungsabsicht mehr beeinflusste als die Wahrnehmung der Angemessenheit des Testverfahrens. Bei Teilnehmern, welche sich in der Mitte der Karriereleiter befinden, spielt dann die Wahrnehmung der Angemessenheit der Testverfahren eine viel grössere Rolle bei der Beeinflussung der Kündigungsabsichten.

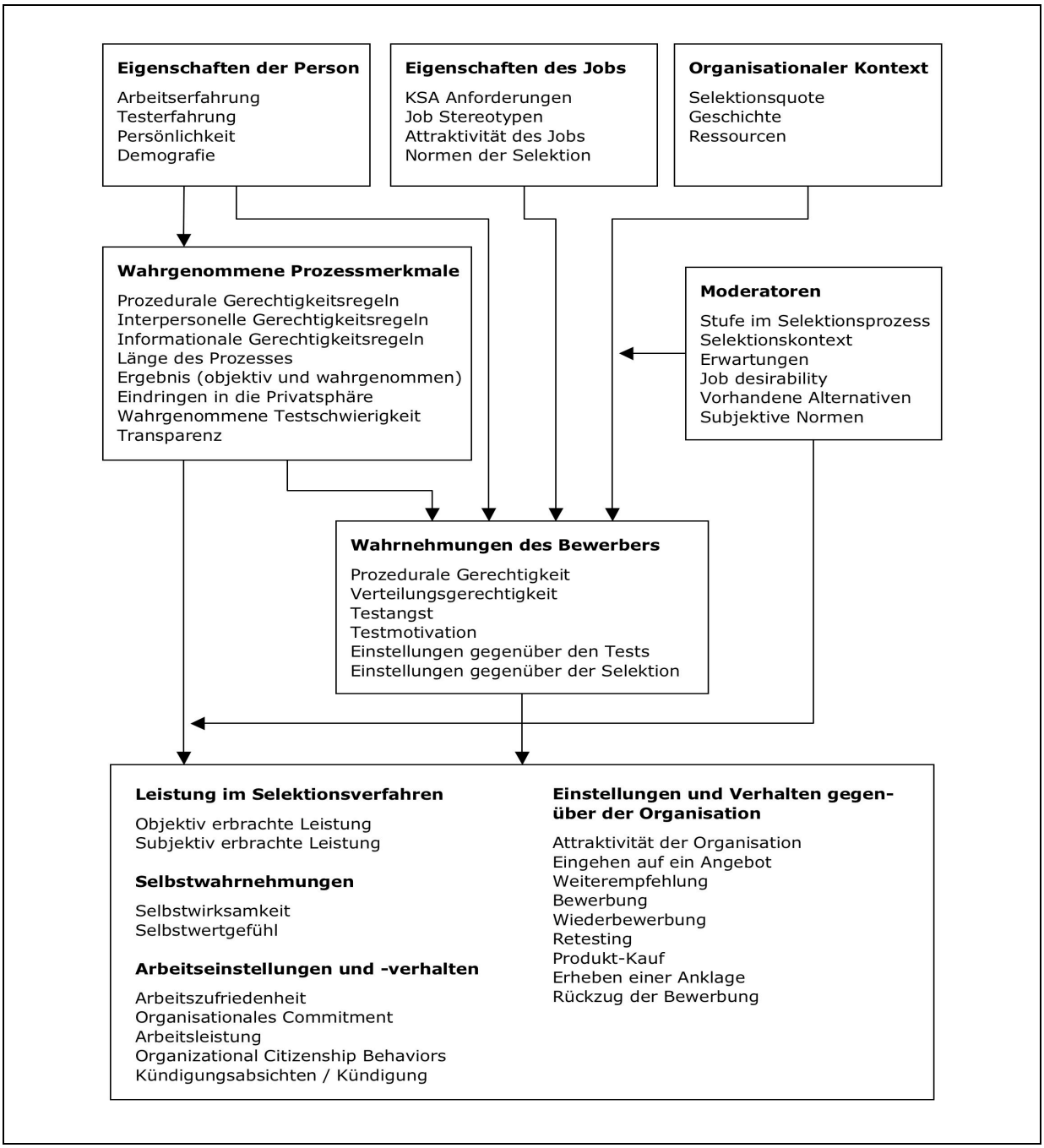


Abbildung 5.2 Das integrative Modell der Bewerberreaktionen auf Personalauswahlverfahren von Hausknecht et al. (2004).

5.5.2 Struktur der Bewerberreaktionen im Militär von Schreurs

Das Ziel der Studie von Schreurs (2007; siehe auch Derous & Schreurs, 2009) war die Entwicklung eines kontentspezifischen Modells der Bewerberreaktionen auf den Selektionsprozess im Belgischen Militär, welcher folgende drei Stufen umfasst: Information im Laufbahnbüro, Intelligenz-Screening und ausführliches Screening (Medizin, Sport, Persönlichkeit). Damit folgten sie der Empfehlung von Chan und Schmitt (2004) und erstellten ein Modell, welches genau die interessierende Gegebenheit abbildet – das Modell umfasst zum Beispiel sehr spezifische Aspekte wie die Wahl der Einsatzverwendung, Erreichbarkeit und Unterkunft oder die Unzufriedenheit mit Wartezeiten – und damit im Gegensatz zur Forschungstradition von Gilliland (1993) steht, bei welcher allgemeingültige Aspekte der Bewerberreaktionen untersucht werden.

Dabei ging Schreurs in zwei Phasen vor: Auf der Basis der Grounded Theory (Glaser & Strauss, 1967; Strauss & Corbin, 1997) entwickelte er in drei Schritten die erste Version des Modells zu Bewerberreaktionen. Zuerst führte er 250 Interviews mit ehemaligen Bewerbern auf einen Posten in der Belgischen Armee zu deren Erfahrungen während der drei Selektionsstufen durch. Die transskribierten Inhalte unterzog er einer Inhaltsanalyse und ordnete sie vordefinierten, aus der Literatur (z. B. Bauer et al., 2001; Derous et al., 2004; Gilliland, 1993) gewonnenen Kategorien zu. Für nicht zuordenbare Aussagen bildete er neue Kategorien und formulierte zu allen Kategorien abschliessend auf der Grundlage der Interviewaussagen Items. Als Überprüfung der Skalen ordneten Experten (zwei Akademiker und zwei Selektionsoffiziere) im letzten Schritt die Items den gebildeten Kategorien zu. Bei einer Übereinstimmung von drei Experten beliess er das Item in der Kategorie. Das Ausgangsmodell bestand aus sieben Kategorien und 58 Items zur Selektionsstufe „Beratung im Laufbahnbüro", sieben Kategorien und 39 Items zum Intelligenz-Screening und 19 Kategorien und 124 Items zur letzten Stufe, dem ausführlichen Screening.

In der zweiten Phase überprüfte Schreurs das Modell anhand einer Expertenanalyse: 53 Recruiter der Belgischen Armee mussten je 221 auf Karten gedruckte Items anhand der Q-Sort-Technik (Stephenson, 1953) zu Gruppen mit ähnlichem Inhalt sortieren. Die Experten waren frei, die Anzahl unabhängiger Gruppen zu wählen. Abschliessend mussten sie jeder Gruppe eine Überschrift geben. Diese Daten analysierte Schreurs mit einer multidimensionalen Skalierung, welche die ursprünglich gewählten Dimensionen bestätigte. Die MDS-Cluster konnte er zusätzlich mit einer Tree-Modellierung bestätigen. Aufgrund dieser Analyse schloss er insgesamt 63 Items aus. In Tabelle 5.4 habe ich eine

Übersicht über die Kategorien der Bewerberreaktionen in den drei Selektionsstufen in der Belgischen Armee dargestellt.

Im Gegensatz zu bisherigen Forschungsresultaten, welche von einer hohen Standardisierung des Selektionsprozesses als Fairnessmerkmal ausgehen, zeigte sich im Selektionsprozess der Belgischen Armee, dass die Bewerber massgeschneidertes Vorgehen und unterschiedliche Behandlung schätzen, vor allem wenn es um Jobinformationen, die Informationen zum Selektionsprozess, das Feedback zur Testung und um Wartezeiten im Prozess geht.

Tabelle 5.4

Übersicht über die Kategorien der Bewerberreaktionen in den drei Selektionsstufen in der Belgischen Armee (nach Derous & Schreurs, 2009, S. 53)

Beratung im Laufbahnbüro	Intelligenz-Screening	Ausführliches psychologisches Screening
—	• Augenscheinvalidität	• Augenscheinvalidität • Unvoreingenommene Untersuchung: Diskrimination
• Partizipation: Möglichkeit, Fragen zu stellen	—	—
—	—	• Möglichkeit, sein Potenzial zu zeigen • Faking • Voice
• Dokumentation	• Informationen zum Selektionsprozess: Klarheit der Instruktionen	• Voraus-Information —
• Menge der Information	—	—
• Klarheit der Information	—	—
• Realitätsbezug der Jobinformation	—	
• Professionalität	—	• Professionalität
—	• Rückmeldung	• Rückmeldung
• gute Behandlung	—	• Respekt für den Bewerber
	—	• Faire Behandlung
—	• Behandlung: ruhige Testumgebung	—
—	—	• Eindringen in die Privatsphäre
—	• Bedienungskomfort des Computers	
—	• Wahrgenommener Schwierigkeitsgrad des Tests	• Schwierigkeit der Testung, Höhe der Anforderungen
—	• Wahrgenommener Zeitdruck	• Organisation: Unzufriedenheit mit Wartezeiten und Verzögerungen
—	—	• Organisation: Zufriedenheit mit dem Ablauf
—	—	• Begrüssung des Bewerbers
—	—	• Medizinisch-körperliche Untersuchung
• Wahl der Tätigkeit	—	
• Zugänglichkeit und Unterkunft		—
		—

5.6 Skalen zur Erfassung der Akzeptanz von Testverfahren

Bei einer bewerberzentrierten Personalselektion (Boss, 2005) holt der Personalverantwortliche jeweils nach der Testdurchführung und nach dem Feedbackgespräch eine Rückmeldung zum Erleben der Selektionssituation ein, idealerweise auch von den abgelehnten Bewerbern (Schleicher et al., 2006). Dies gibt dem Personalverantwortlichen einerseits die Möglichkeit, negative Gefühle des Bewerbers aufzufangen und zu besprechen und andererseits erhält er auf diese Weise wichtige Hinweise für die Weiterentwicklung des Selektionsverfahrens. In den von der Society for Industrial and Organizational Psychology (2003) herausgegebenen *Principles for the Validation and Use of Personnel Selection Procedures* nehmen die Autoren im Kapitel über die Akzeptanz von Testverfahren durch den Bewerber auch die Testentwickler in die Pflicht: „In addition to the organization's needs and objectives, researchers also need to consider the acceptability of the selection procedure to candidates ... [and] should consider approaches designed to minimize any negative perception of a selection procedure" (S. 40). Um jedoch die Akzeptanz eines Testverfahrens festzustellen, müssen zuerst entsprechende Erhebungsinstrumente entwickelt werden. In ihrem Übersichtsartikel zur Messung von Bewerberreaktionen auf Selektionsverfahren legen Bauer, Truxillo und Paronto (2003) eine ausführlich dokumentierte Liste mit zwischen 1987 und 2001 publizierten Skalen vor. Da die Autoren zu jeder Skala auch die entsprechenden Items aufführen, stellt ihr Beitrag eine Fundgrube für Testentwickler und für weiterführende Forschungsarbeiten im Bereich der Bewerberreaktionen dar. Die von Bauer et al. (2003) gewählte Gliederung der einzelnen Skalen und Subskalen zeigt zudem die Breite auf, mit welcher Forscher die kurz-, mittel- und langfristigen Auswirkungen von Selektionsverfahren untersuchen: Verfahrensgerechtigkeit, Augenscheinvalidität, prädiktive Validität, Fairness der Entscheide, Vertrauen in die Testverfahren, Zufriedenheit mit dem Selektionsprozess, Attraktivität der Organisation, Kaufabsichten, Verhältnis zwischen den Angestellten, Kündigungsabsichten, Absicht auf eine Klage, Absicht auf eine Wiederbewerbung, Empfehlungsabsichten, Selbstwirksamkeit und Testmotivation.

Nachfolgend stelle ich einige dieser Skalen ausführlich dar, wobei ich mich auf solche zur Erfassung der Akzeptanz von Testverfahren beschränke. Die Auswahl soll einen Überblick über die historische Entwicklung dieses Forschungsgebietes geben und die Bandbreite der methodischen Zugänge aufzeigen.

Skala zur Messung der Reaktion auf Testverfahren (Fiske, 1967)

In einer der ersten Studien zur Akzeptanz von Testverfahren führte Fiske (1967) mit knapp 600 Probanden nach der Bearbeitung von jeweils zwei aus sechs Testverfahren in zwei unterschiedlichen experimentellen Situationen – Bewerbungs- und Forschungssituation – ein standardisiertes Interview durch. Dabei erhob er die Vorerfahrung mit psychologischen Testverfahren (Vorwissen und allgemeine Haltung), die Augenscheinvalidität der sechs eingesetzten Testverfahren und die emotionalen Reaktionen auf die Testdurchführung. Neben der Möglichkeit, die bei der Testbearbeitung auftretenden Emotionen frei zu äussern, legte Fiske den Probanden auch eine Liste mit 13 Emotionen vor – Interesse, Neugier, Missbehagen, Gleichgültigkeit, Erheiterung, Ängstlichkeit, Verärgerung, Langeweile, Sinnlosigkeit, Frustration, Anspannung und Zufriedenheit –, welche er aufgrund der Analyse der freien Äusserungen in einem Pilotversuch erstellt hatte.

Als wichtigste Erkenntnis seiner Studie nennt Fiske (1967) die Vielfalt unterschiedlicher Reaktionen auf die Tests und auf das Getestetwerden und zieht daraus folgenden Schluss:

> The study suggests that, after a test, it is often desirable to assess the subjects' interpretations of the purposes of the test, their evaluations of it, and their feelings about it. ... Anyone administering and interpreting tests must recognize that subjects are not completely in the dark about the purposes of the test and about their level of utility, and that subjects do react to being tested. (S. 294-295)

Fiske regte an, dass auf mögliche Reaktionen auf Testverfahren schon bei der Testentwicklung zu achten ist, nicht etwa um den Test und die Testung möglichst angenehm zu gestalten, sondern um Störeinflüsse auf die Testergebnisse zu erkennen und zu reduzieren. Dass es Fiske nicht darum ging, die Ergebnisse aus der Befragung von Testbearbeitern als Grundlage für eine Verbesserung der Testsituation zu nutzen, ist insofern erstaunlich, als in den Vereinigten Staaten von Amerika schon in den sechziger Jahren eine rege Debatte über den Einsatz psychologischer Testverfahren geführt wurde. Die Testentwickler reagierten darauf aber – wie dieses Beispiel zeigt – mit einer vertieften Analyse der psychometrischen Grundlagen von Testverfahren und rückten den Testbearbeiter erst Jahre später in den Fokus ihrer Forschung, indem sie sich mit den Aspekten der diagnostischen Situation und der Fragestellung, wie sie diese angenehmer gestalten können befassten.

Examinee Feedback Questionnaire (EFeQ; Nevo & Sfez, 1985)

Nevo und Sfez (1985; siehe auch Nevo, 1993, 1995) entwickelten den *Examinee Feedback Questionnaire* (EFeQ) für den Einsatz bei der interuniversitären psychometrischen Eintrittsprüfung in Israel, welche aus fünf Testverfahren besteht (Allgemeinwissen, figurales Schlussfolgern, Textverständnis, mathematisches Schlussfolgern und Englisch). Die Autoren gingen dabei von ethischen Überlegungen aus, indem sie mit ihrer Befragung den angehenden Studenten zeigen wollten, dass man sich für ihre Haltung gegenüber den eingesetzten Testverfahren und ihr Erleben der Testung interessiert. Testverfahren oder Selektionsprozesse, welche von den Bewerbern gut aufgenommen werden, können geringere Messfehler aufweisen, kompetente Bewerber anziehen, Unzufriedenheit bei abgewiesenen Bewerbern senken und zu einem besseren Ansehen in der Öffentlichkeit führen (Nevo & Sfez, 1985). Für die Prozessverantwortlichen bietet der EFeQ auch eine Möglichkeit zur Überwachung der für die Testdurchführung erlassenen Vorschriften und als Grundlage für die Weiterentwicklung der Testverfahren. Zudem werteten Forschergruppen die auf diese Weise erhobenen Daten im Hinblick auf Fragestellungen im Bereich der Sozialpsychologie, der interkulturellen Psychologie und zur Testängstlichkeit aus. Für die Entwicklung des EFeQ haben die Autoren die dazumal spärlich vorhandene Literatur zu dieser Thematik konsultiert und anschliessend Items entwickelt, welche Themen ansprechen, welche Testentwickler und/oder Testbearbeiter als wichtig erachten.

Der EFeQ erhebt das Verhalten, die Einstellungen und die Gefühle der Getesteten zum Testverfahren, den Testleitern und der Testsituation vor, während und nach der Testung. Um sozial erwünschtes Antworten auszuschliessen, wird er anonym ausgefüllt. Der Fragebogen besteht aus drei Teilen mit insgesamt sieben bis acht Fragen zu im Zusammenhang mit der Testung relevanten Themen (Nevo & Sfez, 1985):

1. Allgemeiner Teil zu den Testbedingungen und dem Verhalten der Testleiter.
2. Fragen zur Testsituation (Testinstruktionen, Antwortblätter, Testumgebung, physikalische Umstände bei der Testung), zum Test (Augenscheinvalidität, Fairness, Attraktivität, Testdauer, Schwierigkeit) und zum Testbearbeiter (Testvorbereitung, Selbsteinschätzung der Testleistung, emotionale Testreaktion, Vorerfahrung).
3. Eine offene Frage, welche die Möglichkeit bietet, Kommentare abzugeben.

Die Befragung wird für alle Testverfahren am Schluss der Testbatterie durchgeführt und dauert 15 bis 20 Minuten. Zu den einzelnen Fragen stehen den Testbearbeitern jeweils fünf bis sechs abgestufte Antworten zur Verfügung. In der nachfolgenden Liste tragen sie für jeden bearbeiteten Test die Nummer der gewählten Antwortalternative ein. Die EFeQ-Konstrukteure tauschen die Fragen von Zeit zu Zeit aus, um jeweils zu einer spezifischen Thematik die Meinungen einzuholen. Weiter schlagen sie vor, dass der Fragebogen für jeden Verwendungszweck anzupassen ist.

Nevo (1993) sieht in diesem Fragebogen auch die Chance für unzufriedene Testnehmer, ihrem Ungerechtigkeitsempfinden Ausdruck zu verleihen und so ihr Stresserleben zu senken. Ebenso ist er der Ansicht, dass diese Art der Erhebung bei jedem Testverfahren durchgeführt werden sollte und die Ergebnisse daraus im Testmanual aufzuführen sind. Da das Fairnesserleben jedoch stark von der jeweiligen Testsituation abhängig ist, sind generalisierte Aussagen dazu nicht möglich. Der Schwachpunkt des EFeQ liegt in der fehlenden theoretischen Untermauerung. Nevo (1993) empfiehlt deshalb in seinem Leitfaden für die Entwicklung von auf die jeweilige Testsituation abgestimmte EFeQs, dass als Grundlage ein Literaturstudium zu den Haltungen gegenüber Testverfahren, Selbstreports von Getesteten und Berichten über andere Feedback-Fragebogen durchzuführen sei und in einem zweiten Schritt die Feedback-Bedürfnisse der Organisation geklärt werden müssen.

In der Zeit von 1987 bis 1993 nimmt das Interesse an der Erforschung der Reaktionen von Bewerbern auf Testverfahren deutlich zu und mehrere Forschergruppen publizieren Skalen, welche sich jedoch auch nicht auf eine theoretische Basis abstützen. Dazu zählen zum Beispiel die aus zehn Items bestehende Skala von Ryan und Sackett (1987) zur Erfassung der Einstellung zu Testverfahren, mit Fragen zur Verbreitung des Einsatzes von Testverfahren in der Selektion, zur Bereitschaft, den Test zu bearbeiten, zur Verletzung der Privatsphäre oder zum Bild der Organisation, welche solche Tests einsetzt. Lounsbury, Bobrow und Jensen publizierten 1989 ihre aus 25 Items bestehende Skala zur Erfassung der Einstellungen gegenüber Selektionsverfahren, welche sie auf der Basis von publizierten Studien (u. a. Fiske, 1967; Tesser & Leidy, 1968) entwickelt haben. Mittels einer Faktorenanalyse bestimmten sie einen Faktor, auf welchen 17 Items zu den Aspekten Validität, Fairness, Augenscheinvalidität, Nützlichkeit und Verletzen der Privatsphäre laden. Weitere Beispiele aus dieser Zeit sind die Skala zur Erfassung der Bewerberreaktionen bei Selektionsverfahren von Kluger und Rothstein (1993), welche aus 28, theorielos zusammengestellten Dimensionen, wie

antizipierte Jobperformance, Image der Unternehmung, Relevanz, Testfairness, Schwierigkeit oder Schläfrigkeit besteht oder die Skala von Rynes und Connerley (1993) mit fünf Items zur Angemessenheit des Testverfahrens und zur Haltung gegenüber der Organisation. Iles und Robertson (1989) waren dann unter den ersten Autoren, welche anhand des von ihnen als *impact validity* bezeichneten Konzeptes ein Kausalmodell der Bewerberreaktion entwickelten. Sie unterscheiden darin vier Gruppen von Variablen: Unabhängige Variablen (Testmethode und Selektionsentscheidung), mediierende Variablen (Bewerberreaktionen, Überzeugungen und Einstellungen), Moderatorvariablen (Persönlichkeit und berufliche Situation) und Outcome-Variablen (Selbstwertgefühl, Commitment, Kündigungsabsichten). Eine eigenständige Skala zur Erfassung der im Modell aufgeführten Bewerberreaktionen entwickelten sie jedoch nicht.

Test Attitude Survey (TAS; Arvey, Strickland, Drauden & Martin, 1990)

Ausgangspunkt für die Entwicklung der *Test Attitude Survey* (TAS) ist die Bedeutung der Testmotivation bei prädiktiven und konkurrenten Validierungsstudien von Testverfahren: Vergleichen die Testkonstrukteure bei einer prädiktiven Validierungsstudie die Testleistung von Bewerbern mit deren später gezeigten Leistung im Beruf, legen sie häufig zur Bestimmung der konkurrenten Validität aktuellen Jobinhabern die zu überprüfende Testbatterie vor und vergleichen die Ergebnisse mit der Berufsleistung. Es ist offensichtlich, dass sich das Engagement und die emotionale Belastung bei der Testbearbeitung bei den Bewerbern und den Jobinhabern deutlich unterscheiden und dies einen Einfluss auf die jeweiligen Validitätskoeffizienten haben kann. Arvey et al. (1990) interessierten sich zudem für die Fragen, ob testbezogene Motivation und Einstellungen einen Einfluss auf die Testleistung haben, ob diesbezügliche Unterschiede zwischen verschiedenen Personengruppen existieren und ob die Validität eines Testverfahrens dadurch beeinflusst wird. Um Antworten auf ihre Fragen zu erhalten, entwickelten sie die TAS.

Dazu wendeten die Autoren eine rationale Testkonstruktion kombiniert mit empirischer Evidenz an: In einem ersten Schritt entwickelten sie Fragen zu grundlegenden Konstrukten der Testmotivation wie Leistungsmotivation, Herausforderung durch den Test, Schwierigkeit des Tests oder subjektiv eingeschätzte Auswirkungen des Testergebnisses. Eine erste Pilotstudie zeigte auf, dass der gewählte Motivationsbegriff zu eng war und so erweiterten die Autoren ihren Fragebogen um Aspekte wie Zuschreibung des Testerfolges, Wahrnehmung des Zeitdrucks und Augenscheinvalidität. Nachdem eine erste Validierungsstudie, in welcher Experten die 60 Fragen vordefinierten Kategorien zuordnen mussten,

gute Ergebnisse lieferte, führten die Autoren mittels Daten aus realen Bewerbungssituationen eine Faktorenanalyse durch, welche sieben Faktoren mit einem Eigenwert grösser als eins ergab. Da sich diese Faktoren mit den a priori festgelegten Faktoren nicht genügend deckten, nahmen die Autoren die Resultate aus der Experteneinstufung hinzu und entwickelten auf dieser Grundlage und ihrem eigenen Fachwissen die definitive Version des Fragebogens, welche die nachfolgend aufgeführten neun Dimensionen enthält (in Klammern sind jeweils die Anzahl der Items angegeben): Testmotivation (10), Konzentrationsmangel (4), Vertrauen in die Testverfahren (4), Testangst (10), Einfachheit des Tests (4), externale Attribution der Testleistung (5), Leistungsmotivation (3), Auswirkungen des Testergebnisses für die Zukunft (3), Vorbereitung (2).

Die Reliabilitäten der Skalen liegen zwischen α =.56 und .85. Arvey et al. (1990) belegten die Validität der TAS, indem sie die Einstufungen aus einem schwierigen Test mit denjenigen eines einfachen Tests und diejenigen eines computergestützten Verfahrens mit denjenigen einer Papier-und-Bleistift-Version verglichen. In beiden Studien konnten sie die erwarteten Effekte nachweisen. Es zeigte sich zudem, dass sich die Einstufungen der Bewerber und Jobinhaber in allen Skalen der TAS mit Ausnahme des Vertrauens in die Testverfahren und der Einfachheit der Tests unterscheiden, insbesondere dass sich erstere bei der Testdurchführung deutlich mehr anstrengen als letztere. Damit konnten die Autoren belegen, dass sich die Testleistung aus der im Test gemessenen Fähigkeit, einem Messfehler und der Testmotivation zusammensetzt.

Arvey et al. (1990) legten mit ihrer *Test Attitude Survey* eine nach wissenschaftlichen Kriterien entwickelte Skala mit akzeptablen Testkennwerten vor, welche sich – in Ermangelung besserer Alternativen – an Theorien der Motivationspsychologie anlehnt. Es stellt sich aus diesem Grund jedoch die berechtigte Frage, ob die neun Dimensionen der TAS die wichtigsten Faktoren der Akzeptanzbeurteilung von Testverfahren abdecken. Dies zeigte sich auch in später durchgeführten Überprüfungen der Skala: In einer Studie von Schmit und Ryan (1992) ergab sich eine Einfaktorenstruktur und die von McCarthy und Goffin (2003) durchgeführte explorative Faktorenanalyse ergab eine Dreifaktorenstruktur mit den Dimensionen Testmotivation, Selbstzweifel bezüglich der Testung und Ablehnung von Tests.

Bewerberreaktionen bei Selektionsverfahren (Smither, Reilly, Millsap, Pearlman & Stoffey, 1993)

Smither et al. (1993) messen den Reaktionen der Bewerber bei Selektionsverfahren einen hohen Stellenwert bei, da diese zum Teil weit reichende Konsequenzen haben können:

1. Attraktivität der Organisation: Die Arbeitgeber sollten wissen, welche Testverfahren von den Bewerbern bevorzugt werden und welche auf Ablehnung stossen. Der Einsatz unbeliebter Testverfahren kann dazu führen, dass sich die Bewerber ein schlechtes Bild der Organisation machen.
2. Ethische und juristische Auswirkungen: Bewerber empfinden Testverfahren, welche unvalide erscheinen oder zu tief in die Privatsphäre eindringen, als unfair, unethisch oder unmoralisch, was zu vermehrten Beschwerdefällen führen kann.
3. Reliabilität und Validität der Testverfahren: Nehmen Bewerber Testverfahren als unfair oder unvalide wahr, so kann dies die Leistungsmotivation senken und zu einer schlechteren Testleistung führen.

Für die Entwicklung eines Messverfahrens zur Erfassung der Bewerberreaktionen schlagen Smither et al. ein theoriegeleitetes Vorgehen vor, wobei sie jedoch beklagten, dass diesbezüglich wenig theoretische Vorarbeit geleistet wurde. Bei ihrem Fragebogen stützten sie sich auf das von Herriot (1989) für die sozialpsychologische Theorie des Selektionsprozesses vorgeschlagene Modell der Rollenbildung von Katz und Kahn (1978), die Theorie der sozialen Validität von Schuler (1993a) und das von Stoffey et al. (1991) zur Erklärung der Bewerberreaktionen verwendete Modell der Gerechtigkeitsforschung. Der Fragebogen umfasst 23 Items und sechs Skalen, welche stark an das Schuler-Modell angelehnt sind, wie Tabelle 5.5 zu entnehmen ist.

Smither et al. zeigten mit ihrem Fragebogen unter anderem auf, dass sich die Wahrnehmung der prädiktiven Validität (Zusammenhang zwischen der Testleistung mit der Arbeitsleistung) von der Augenscheinvalidität (Zusammenhang des Testinhaltes mit dem Arbeitsinhalt) unterscheiden lässt. Mit ihrem Fragebogen zu Bewerberreaktionen gelang es Smither et al. empirische Evidenz für das Modell von Schuler (1993a) zu erbringen und sie legten damit das erste Messinstrument vor, das sich an einem Modell der Auswirkungen von Selektionssituationen auf Bewerber orientiert.

Tabelle 5.5
Vergleich der Skalen von Smither et al. (1993) mit dem Modell der sozialen Validität von Schuler (1993a)

Smither, Reilly, Millsap, Pearlman und Stoffey (1993)	Schuler (1993a)
Augenscheinvalidität	Transparenz
Wahrgenommene prädiktive Validität	Transparenz
Wahrscheinlichkeit der Testleistungsverbesserung	Partizipation
Emotionale Reaktion auf die Testung	–
Wissen über das eigene Abschneiden im Test	Urteilskommunikation
Attraktivität der Organisation	–
–	Information

Selection Fairness Survey (Gilliland & Honig, 1994)

Mit der Selection Fairness Survey (SFS) haben Gilliland und Honig (1994) ein Messinstrument zur Erfassung des Gilliland-Modells der Bewerberreaktionen (1993) entwickelt, welches in der Endversion elf Unterskalen und insgesamt 40 Items umfasst. Ziel war es, für eine weitere Erforschung der Gerechtigkeit in Selektionsprozessen ein Instrument zur Verfügung zu stellen, welches die Bewerberreaktionen anhand der prozeduralen und distributiven Gerechtigkeitsregeln erfasst. Die zu den zehn Gilliland-Regeln zur prozeduralen Gerechtigkeit und den drei Verteilungsgerechtigkeitsregeln (*equity*, *equality* und *needs*) entwickelten Items stammen einerseits aus Publikationen zu Bewerberreaktionen (Bies & Shapiro, 1988; Folger & Konovsky, 1989; Kluger & Rothstein, 1993; Konovsky & Cropanzano, 1991; Lounsbury et al., 1989; Schmitt & Coyle, 1976; Smither et al., 1993) und andererseits aus Critical Incidents-Interviews über die Fairness von Selektionsverfahren mit Arbeitsplatzsuchenden (Gilliland, 1995). Experten ordneten die 290 gesammelten Incidents den zehn Gilliland-Regeln zu. Dabei ging die Dimension „Möglichkeit zur Wiedererwägung" leer aus. Hingegen ergab sich eine zusätzliche Kategorie „Einfachheit der Verfälschung" (*ease of faking*), zu welcher sich keine Anhaltspunkte in der Literatur zur Gerechtigkeitsforschung finden liessen. Die erste Version der SFS bestand aus 81 Items zu neun der zehn Gilliland-Faktoren, den drei distributiven Regeln und der neuen Kategorie. Eine erneute Zuordnung der Items zu diesen 13 Dimensionen durch Experten führte dazu, dass die Testentwickler die Equality-Regel weglassen mussten, da sich diese mit anderen prozeduralen Regeln – Möglichkeit zur Selbstdarstellung und Vergleichbarkeit der Durchführung – überschnitt. Die mit

den verbleibenden 56, fünfstufig likert-skalierten Items durchgeführte Datenerhebung und anschliessende Skalenanalyse führte zum Ausschluss der Skalen „Vergleichbarkeit der Durchführung" und „needs". Dies macht durchaus Sinn, da Bewerber selten beurteilen können, ob das Selektionsverfahren für alle Teilnehmer vergleichbar ist. Zudem bildeten Gilliland und Honig aus einem Teil der Items der Skala „Angemessenheit der Fragen" eine weitere Skala „unvoreingenommene Beurteilung" (*consistency bias*).

Das definitive Messinstrument besteht somit aus folgenden elf Dimensionen: Tätigkeitsbezug, Möglichkeit zur Selbstdarstellung, Ergebnisrückmeldung, Information zum Auswahlverfahren, Aufrichtigkeit, respektvolle Behandlung, Zweiweg-Kommunikation, Angemessenheit der Fragen, unvoreingenommene Beurteilung, Einfachheit der Verfälschung, faire Beurteilung (equity). Im Gegensatz zur *Test Attitude Survey* (TAS, Arvey et al., 1990), welcher die Motivation der Testbearbeiter erfasst, erhebt die SFS also Fairnessreaktionen. Gilliland und Honig weisen darauf hin, dass man nicht für jede Fragestellung und jede Ausgangslage alle elf Dimensionen sinnvoll einsetzen kann und das Messinstrument dementsprechend anzupassen ist. Trotzdem konnte es sich nicht durchsetzen und auf Grund der eher mässigen Homogenität der Subskalen raten Bauer, Truxillo und Paronto (2003) sogar von dessen Verwendung ab. In einer überarbeiteten Version der SFS konnten Gilliland und Beckstein (1996) diesen Mangel jedoch beheben.

Selection Procedural Justice Scale (SPJS; Bauer, Truxillo, Sanchez, Craig, Ferrara & Campion, 2001)

Mit der SPJS haben Bauer et al. (2001) eine reliable und valide Skala zur Messung der Verfahrensgerechtigkeitsregeln von Gilliland (1993) entwickelt. Damit erreichten sie, dass Forscher in Zukunft nicht mehr für jede Studie zu den Bewerberreaktionen eine neue Skala entwickeln müssen, da diese „ad hoc-Messinstrumente" die Gefahr bergen, die Forschung zu fragmentieren (Heneman, 1985). In folgenden fünf Schritten entwickelten sie ihre Skala:

1. Entwicklung der Items

Auf der Basis der Definitionen zu den zehn Verfahrensgerechtigkeitsdimensionen von Gilliland und den Items von Bauer et al. (1998) zu fünf Gilliland-Faktoren entwickelten die Autoren einen 50 Items umfassenden Pool, welchen sie von fünf Experten und Personalverantwortlichen hinsichtlich der Inhaltsvalidität und des Wordings überprüfen liessen.

2. Reduktion des Itempools
Anhand der Daten von 330 Bewerbern führten Bauer et al. eine Reliabilitätsanalyse (Cronbach Alpha zwischen .73 und .92) und eine explorative schiefwinklig rotierte Faktorenanalyse durch, welche zu elf Faktoren führte, wobei der Tätigkeitsbezug in die Inhalts- und prädiktive Validität zerfiel. Eine Faktorenanalyse zweiter Ordnung ergab in Übereinstimmung mit der Theorie von Greenberg (1993a) einen Faktor mit sozialen Aspekten (Vergleichbarkeit der Durchführung, Aufrichtigkeit, respektvolle Behandlung, Zweiweg-Kommunikation und Angemessenheit der Fragen) und einen mit strukturellen Aspekten (Tätigkeitsbezug (prädiktiv), Information zum Auswahlverfahren, Möglichkeit zur Selbstdarstellung, Möglichkeit zur Wiedererwägung und Ergebnisrückmeldung). Die Dimension Tätigkeitsbezug (Inhalt) ergab den dritten Faktor.

3. Überprüfung der Faktorenstruktur
Eine konfirmatorische Faktorenanalyse mit den trennschärfsten 39 Items und den Daten von 242 Bewerbern und Trainees ergab den besten Modell-Fit für das Elf-Faktoren-Modell.

4. Überprüfung der Konstruktvalidität
Anhand einer Gruppe von 70 Trainees konnten Bauer et al. nachweisen, dass der SPJS die Verfahrensgerechtigkeit erfasst. Zudem zeigten sich signifikante Korrelationen mit der Attraktivität der Organisation, dem Commitment zur Organisation, Empfehlungsabsichten, dem Selbstwert und der Verteilungsgerechtigkeit.

5. Replikation und Verallgemeinerung
Mit zwei je ca. 200 Studenten umfassenden Stichproben konnten sie die Faktorenstruktur und die konvergente und diskriminante Validität bestätigen.

Es zeigte sich, dass die drei Faktoren zweiter Ordnung (soziale und strukturelle Aspekte und tätigkeitsbezogener Inhalt) die besten Masse für die Verfahrensgerechtigkeitsregeln sind. Insgesamt haben sich die elf Faktoren jedoch auch bewährt und können somit ebenfalls für die Erforschung der Bewerberreaktionen eingesetzt werden. Fazit: Der SPJS stellt ein seriös und äusserst konsequent entwickeltes und überprüftes Verfahren zur Messung des Gilliland-Modells dar.

Social Process Questionnaire on Selection (SPQS, Derous, Born & De Witte, 2004)

Grundlage des SPQS bildet das *Social Process Model on Selection* (SPS model) von Derous und De Witte (2001; siehe auch Derous, De Witte & Stroobants, 2003), welches sich seinerseits an den Modellen von Schuler (1993a) und Gilliland (1993) anlehnt. Die in Tabelle 5.6 dargestellten Abweichungen zu den

beiden Modellen ergaben sich auf Grund von Gesprächen der Autoren mit Bewerbern und Personalverantwortlichen (Derous & De Witte, 2001). Zusätzlich unterscheidet sich das SPS-Modell von den beiden anderen Modellen, indem es sich nicht nur auf Reaktionen der Bewerber nach einer Testung bezieht, sondern auch Pretest-Reaktionen erfasst, welche wichtige Einflussfaktoren auf die Testleistung und die Posttest-Reaktionen darstellen (z. B. Chan, Schmitt, Sacco & DeShon, 1998).

Zu den einzelnen SPS-Dimensionen formulierten Derous et al. (2003) insgesamt 69 Items, welche sie einer Analyse durch 30 Recruiting- und Selektions-Experten unterzogen. Dabei bekamen die Experten die Aufgabe, die Items in einem von ihnen selbst definierten Kategoriensystem zu gruppieren (Q-Sort-Technik; Stephenson, 1953). Eine mit diesen Q-Sorts durchgeführte multidimensionale Skalierung ergab ein Modell mit sechs Regionen (Information zur Arbeitsstelle, Partizipation, Transparenz, Feedback, Objektivität und humane Behandlung) entlang von zwei Dimensionen (Differenzierung vs. Standardisierung und Aufgabe vs. Beziehung). Die erste Version des anhand dieses Modells konstruierten *Social Process Questionnaire on Selection* umfasste insgesamt 46 Fragen und wurde in zwei Versuchsbedingungen in einer Pretest-Situation überprüft: Die Probanden der einen Stichprobe mussten angeben, für wie wichtig sie diese Aussage im Rahmen einer Bewerbungssituation halten, diejenigen der zweiten Stichprobe zusätzlich ihre Erwartungen, ob der im Item beschriebene Sachverhalt in der gleich folgenden Testung auch umgesetzt wird. Die explorative Faktorenanalyse bestätigte die sechs Faktoren, welche gut 50% der Varianz erklärten. Insgesamt 41 Items wiesen Faktorladungen über .40 auf und 37 davon übernahmen die Autoren in die definitive Version des Fragebogens, welchen sie anhand einer weiteren Stichprobe (N = 634) mittels einer konfirmatorischen Faktorenanalyse überprüfen und bestätigen konnten.

Mit dem SPQS stellen Derous et al. (2004) Forschern und Praktikern ein äusserst sorgfältig konstruiertes und an aktuelle Modelle angelehntes Instrument zur Verfügung. Besonders hervorzuheben sind die verschiedenen Validierungsstudien und die Einbettung der einzelnen Dimensionen in ein Modell der gelernten Hilflosigkeit im Rahmen der Personalselektion.

Tabelle 5.6

Vergleich des Social Process Model on Selection und des Social Process Questionnaire on Selection mit den Modellen von Schuler (1993) und Gilliland (1993) (nach Derous & De Witte, 2001)

Schuler (1993) *Soziale Validität*	Gilliland (1995) *Procedural justice rules of selection*	Derous und De Witte (2001) *Social process characteristics of selection*	Derous, Born und De Witte (2004) *Social Process Questionnaire on Selection*
Information	Information zum Auswahlverfahren	Allgemeine Informationen über die zu besetzende Stelle	Allgemeine Informationen über die zu besetzende Stelle
Partizipation	Möglichkeit zur Selbst-darstellung	Zulassen der Kontrolle durch den Kandidaten	Aktive Teilnahme am Auswahl-verfahren
	Möglichkeit zur Wieder-erwägung	Offenheit für ein selbstbewusstes Vorgehen des Kandidaten	
	Zweiweg-Kommunikation		
Transparenz	Tätigkeitsbezug	Schaffen einer transparenten Testung	Transparenz der Testung
Urteilskommunikation	Ergebnisrückmeldung	Geben von Feedback	Geben von Feedback
	Vergleichbarkeit der Durchführung	Garantieren eines objektiven Auswahl-verfahrens	Objektives Auswahlverfahren durch professionelle Durchführung und Gleichbehandlung der Kandidaten
	Angemessenheit der Fragen	Wahrung der Privatsphäre und Tätig-keitsbezug der erhobenen Informationen	Gewährleistung einer menschen-würdigen Behandlung und Wahrung der Privatsphäre
	Respektvolle Behandlung	Gewährleistung einer menschenwürdigen Behandlung	
	Aufrichtigkeit		

Applicant Expectation Survey (AES, Schreurs, Derous, Proost, Notelaers & De Witte, 2008)

Schreurs et al. entwickelten mit der AES eine reliable Skala zur Erfassung der Erwartungen der Bewerber hinsichtlich eines bevorstehenden Selektionsprozesses (*selection expectations*; Bell et al., 2006; Derous et al., 2004). Als Grundlage dienten die von Derous und Schreurs (2009) durchgeführten Studien für die Entwicklung des Modells der Bewerberreaktionen in der Belgischen Armee: Studium der Literatur zu Bewerberreaktionen, Durchführung von 250 Interviews mit Bewerbern und die Überprüfung des Modells durch 53 Recruiter (ausführliche Beschreibung siehe Kapitel 5.5.2). Das daraus abgeleitete Modell der Erwartungen zu Selektionsverfahren umfasst die fünf Dimensionen Wärme/Respekt, Möglichkeit, sein Potenzial zu zeigen, Schwierigkeit des Betrügens, unvoreingenommene Beurteilung und Rückmeldung, welche im AES-Fragebogen mit 26 Items operationalisiert sind.

Die Akzept!-Skalen (Kersting, 2008)

Für die in der Personalselektion häufig verwendeten Verfahrensklassen Leistungstests, Persönlichkeitstests und Assessment Center hat Kersting (2006, 2008) Skalen zur Erfassung der Akzeptanz entwickelt, wobei er die sieben Dimensionen der Akzept!-Skalen für Leistungs- (L) und Persönlichkeitstests (P) hauptsächlich von der Kriterienliste von Gilliland (1993) ableitet, wie nachfolgende Tabelle 5.7 aufzeigt:

Tabelle 5.7
Vergleich der Akzept!-Skalen L und P mit der Kriterienliste von Gilliland (1993)

Akzept!-Skalen L und P	Kriterienliste von Gilliland (1993)
Messqualität (L & P)	Vergleichbarkeit der Durchführung Möglichkeit zur Selbstdarstellung
Augenscheinvalidität (L & P)	Tätigkeitsbezug
Kontrollierbarkeit (L & P)	Zweiweg-Kommunikation
Wahrung der Privatsphäre (P)	Angemessenheit der Fragen
Intention zur unverfälschten Antwort (P)	*Einfachheit des Verfälschens (Gilliland & Honig, 1984)*
Antwortfreiheit (P)	Möglichkeit zur Selbstdarstellung
Belastungsfreiheit (L)	*Faire, nicht belastende und angemessene Situation (Schuler, 1990)*

Anmerkung. L = Leistungstests, P = Persönlichkeitstests

Jede Dimension operationalisiert Kersting durch vier Aussagen, welche als Antwortformat eine sechsstufige Likertskala (Extrempole „trifft nicht zu" resp. „trifft genau zu") aufweisen. Der Akzept!-AC enthält zusätzlich die Dimensionen „gute Organisation" und „positive Atmosphäre". Mit zwei Zusatzitems erhebt der Akzept! noch eine Gesamtbeurteilung des Testverfahrens („Welche Schulnote würden Sie dem soeben bearbeiteten Verfahren geben?") und die Einschätzung der eigenen Leistung im Verfahren.

In mehreren Studien überprüfte Kersting (2008) die Gütekriterien der Akzept!-Skalen. Die Reliabilitäten der Subskalen liegen zwischen α = .65 und .82. Die Retestreliabilität bei einem durchschnittlichen Abstand der beiden Messungen von einem Jahr beträgt r_{tt} = .82. Mittels einer konfirmatorischen Faktorenanalyse konnte Kersting (2008) die Dimensionsstruktur bestätigen, wobei die Subskalen Messqualität und Augenscheinvalidität mit r = .67 stark korrelieren. Kersting belässt diese jedoch trotzdem als unabhängige Dimensionen im Fragebogen, da die Testautoren bei der Entwicklung des mit dem in dieser Studie beurteilten Verfahrens – dem WIT-2 (Kersting, Althoff & Jäger, 2008) – besonders darauf achteten, dass das Aufgabenmaterial aus dem Berufsalltag stammt und somit eine hohe Augenscheinvalidität aufweist. Im Gegensatz zu den Ergebnissen aus der Studie von Arvey et al. (1990) ergaben sich bei den Akzept!-Skala für Leistungstests – mit Ausnahme der Gesamtbeurteilung des Verfahrens – keine signifikanten Unterschiede zwischen den Einschätzungen aus einer Bewerbungssituation und einer Laborstudie.

Die Akzept!-Skalen von Kersting (2008) stellen das erste Messinstrument für die Erhebung der Akzeptanz von Testverfahren im deutschen Sprachraum dar und zeichnen sich durch ihre theoretische Fundierung und die gewissenhafte psychometrische Überprüfung aus.

5.7 Merkmale der Akzeptanz verschiedener Testverfahren

Bei der Darstellung des Modells der Bewerberreaktionen auf Personalauswahlverfahren von Gilliland (1993) zeigte sich, dass einzelne der von ihm postulierten zehn Fairness-Regeln auch für Testverfahren von Bedeutung sind. Konkret handelt es sich um den Tätigkeitsbezug, die Möglichkeit zur Selbstdarstellung und die Angemessenheit der Fragen. Schon bei den in den sechziger Jahren geführten Diskussionen zum Einsatz von psychologischen Tests in der Personalselektion kam klar zum Ausdruck, dass zwischen den verschiedenen Verfahren grosse Unterschiede bezüglich deren Angemessenheit für diesen Einsatzzweck und somit deren Akzeptanz durch die Bewerber bestehen (z. B. Amrine, 1965; Fiske, 1967; Guion, 1967; Simmons, 1968). Bis zu den 90er Jahren des letzten Jahrhunderts war die Bewerberperspektive beim Einsatz von Testverfahren jedoch nur vereinzelt Gegenstand wissenschaftlicher Studien (z. B. Bourgeois, Leim, Slivinski & Grant, 1975; Noe & Steffy, 1987; Schmidt et al., 1977; Teel & Dubois, 1983) und blieb so praktisch unerforscht (Rynes, 1993), obwohl die Notwendigkeit der vertieften Auseinandersetzung mit der Thematik erkannt war (z. B. Thornton & Byham, 1982). Eine Reihe von zu Beginn der 90er Jahre veröffentlichten Artikeln (z. B. Kluger & Rothstein, 1993; Robertson et al., 1991; Rynes, 1993; Rynes & Connerley, 1993; Smither et al., 1993; Stoffey et al., 1991), die Erweiterung des Fairness-Konzeptes um den Aspekt der Bewerberreaktion (Schmitt & Gilliland, 1992) und das Modell von Gilliland (1993) rückten dieses Forschungsgebiet in den Fokus des wissenschaftlichen Diskurses und lösten eine Flut von Publikationen aus. Damit erfüllte sich die 1990 von Greenberg in seinem Übersichtsartikel zur organisationalen Gerechtigkeit geäusserten Prophezeihung:

> The 1990s promise to be a decade in which the viability of organizational justice as a meaningful organizational construct will be fully realized … and will rise [issues of justice and fairness] to the top of the field of organizational behavior's collective research agenda. (S. 425–426)

Eine bezüglich ihres Umfangs und ihrer Bedeutung für nachfolgende Arbeiten beachtenswerte und darum an dieser Stelle ausführlich dargestellte Studie führten Steiner und Gilliland 1996 in den Vereinigten Staaten von Amerika und Frankreich durch, indem sie die Akzeptanz folgender zehn Personalselektionsinstrumente einstufen liessen: Interview, Lebenslauf, Arbeitsprobe, biografischer Fragebogen, Leistungsfähigkeitstest (Intelligenztest), Referenzen und Zeugnisse, Persönlichkeitstest, Integritätstest, persönliche Kontakte (Netzwerk, „Vitamin B") und graphologisches Gutachten. Andere Forschergruppen replizierten diese Studie in den nachfolgenden Jahren in den Ländern Belgien, Deutschland, Frank-

reich, Griechenland, Italien, den Niederlanden, Portugal, Singapur, Spanien und Rumänien (Anderson & Witvliet, 2008; Bertolino & Steiner, 2007; Ispas, Ilie, Iliescu, Johnson & Harris, 2010; Marcus, 2003; Moscoso & Salgado, 2004; Nikolaou & Judge, 2007; Phillips & Gully, 2002; Stinglhamber, Vandenberghe & Brancart, 1999), so dass heute länder- und kulturübergreifende Einstufungen der Akzeptanz der wichtigsten Instrumente der Personalselektion vorliegen.

Steiner und Gilliland befragten Studenten (Durchschnittsalter 20 Jahre) nach der Arbeitsstelle, auf welche sie sich aus heutiger Sicht nach dem Studium bewerben werden. Danach bekamen diese den Auftrag, jedes der zehn Selektionsverfahren im Kontext einer Bewerbung auf ihre präferierte Stelle anhand nachfolgend aufgeführter Fragen und Aussagen auf einer siebenstufigen Antwortskala zu beurteilen:

1. Als wie effektiv würden Sie diese Methode einschätzen, wenn sie dazu eingesetzt wird, qualifizierte Personen für die von Ihnen angegebene Stelle zu finden?
2. Wie würden Sie über die Fairness dieses Vorgehens denken, wenn Sie die Stelle auf Grund dieser Selektionsmethode nicht erhalten würden?
3. Die Methode basiert auf solider wissenschaftlicher Forschung.
4. Diese Vorgehensweise eignet sich zur Identifikation qualifizierter Kandidaten für die zu besetzende Stelle (Augenscheinvalidität).
5. Die Methode erfasst herausragende Qualitäten eines Individuums, welche es von anderen unterscheidet (Möglichkeit, Leistung zu zeigen).
6. Die Selektionsmethode ist unpersönlich und kalt.
7. Die Arbeitgeber sind berechtigt, Informationen anhand dieser Methode von den Bewerbern zu erfassen.
8. Die Methode dringt in die Privatsphäre des Bewerbers ein.
9. Der Einsatz der Methode ist auf Grund deren weiten Verbreitung gerechtfertigt.

In Tabelle 5.8 stelle ich zusammenfassend die Mittelwerte der Datensätze aus den Ländern Frankreich, Griechenland, den Niederlanden, Portugal, Singapur, Spanien, Rumänien und den Vereinigten Staaten von Amerika dar (insgesamt 1'500 Studienteilnehmer) und vergleiche sie mit den Mittelwerten aus der Studie mit deutschen Studenten (Marcus, 2003). Auf der Skala von 1 (am wenigsten bevorzugt) bis 7 (am meisten bevorzugt) erzielen Interviews, Arbeitsproben und der Lebenslauf bei den ersten beiden Fragen nach der Favorisierung einen Durchschnittswert über 5. Werte zwischen 4 und 5 erreichen Intelligenztests,

biografische Fragebogen, Referenzen und Persönlichkeitstests. Tendenziell negativ beurteilt werden die Integritätstests und persönliche Kontakte und deutlich auf Ablehnung stossen graphologische Gutachten, wobei zu beachten gilt, dass diese heute – zumindest im deutschsprachigen Raum – kaum noch zum Einsatz gelangen (Bangerter, König, Blatti & Salvisberg, 2009; Schuler, Hell, Trapmann, Schaar & Boramir, 2007). Wie stark graphologische Gutachten auf Ablehnung stossen, zeigt sich an den Einschätzungen der von Kravitz et al. (1996) befragten amerikanischen Studenten: Diese stuften graphologische Gutachten in etwa als gleich invasiv ein, wie die Erhebung der psychischen Gesundheit oder eine arbeitsmedizinische Untersuchung und sogar invasiver als zum Beispiel Drogentests oder ein Auszug aus dem Strafregister.

Tabelle 5.8
Akzeptanzeinstufung von in der Personalselektion eingesetzten Testverfahren

Selektionsverfahren	Favorisierung (Fragen 1 & 2)	Studie Marcus (2003)	Durchschnitt Fragen 3 – 9	Wissenschaftlich fundiert	Augenscheinvalidität	Möglichkeit zur Selbstdarstellung	Berechtigung des Arbeitgebers	Wahrung der Privatsphäre	Unpersönlichkeit (-)	Weite Verbreitung
Interview	5.37	5.67	5.39	4.29	5.73	5.62	6.03	5.12	5.30	5.62
Arbeitsprobe	5.15	5.34	5.20	4.65	5.60	5.64	5.71	5.28	4.76	4.79
Lebenslauf	5.02	4.85	5.10	4.27	5.48	5.05	5.84	5.18	4.37	5.55
Intelligenztests	4.54	4.10	4.75	4.97	4.91	5.02	5.11	4.93	4.07	4.48
biografischer Fragebogen	4.38	3.20	4.53	4.07	4.62	4.81	5.10	4.58	4.27	4.41
Referenzen und Zeugnisse	4.29	4.91	3.89	3.25	4.48	4.42	5.03	4.53	4.54	4.50
Persönlichkeitstests	4.26	4.18	4.47	4.50	4.56	4.94	4.73	4.31	4.21	4.22
Integritätstests	3.69	3.64	3.94	3.84	3.86	4.07	4.31	4.15	3.96	3.63
persönliche Kontakte	3.13	2.62	3.35	2.17	2.73	2.91	3.43	4.58	4.40	3.42
graphologisches Gutachten	2.20	1.90	2.69	2.77	2.32	2.44	2.73	3.80	3.13	2.10

Anmerkung. N = 1'500, davon n = 836 Studenten, Studie Marcus (2003) N = 213 Studenten.

Die anhand dieser Studien gebildete Reihenfolge der Präferenz der einzelnen Selektionsinstrumente deckt sich auch sehr gut mit derjenigen von Stone-Romero et al. (2003), derjenigen von Kravitz et al. (1996) und den Ergebnissen aus der Studie von Marcus (2003) mit deutschen Studenten. Diese stuften jedoch den

biografischen Fragebogen deutlich schlechter ein (3.20 vs. 4.38), persönliche Kontakte und Intelligenztests leicht schlechter (2.62 vs. 3.13 resp. 4.10 vs. 4.54) und Referenzen besser (4.91 vs. 4.29). Dies lässt sich zum Teil mit der unterschiedlichen Praxis in den verschiedenen Ländern erklären: So werden biografische Fragebogen in Deutschland kaum eingesetzt, wohingegen die Arbeitgeber verpflichtet sind, Arbeitszeugnisse auszustellen, was zum Beispiel in den Vereinigten Staaten von Amerika nicht der Fall ist.

Nachfolgend gehe ich auf die Akzeptanzbeurteilung einiger Selektionsinstrumente noch vertieft ein (siehe auch Cropanzano & Wright, 2003):

Arbeitsproben und Assessment Centers erhielten auch in anderen Studien sehr hohe Fairnesseinstufungen (Bourgeois et al., 1975; Kluger & Rothstein, 1993; Kravitz et al., 1996; Macan et al., 1994; Noe & Steffy, 1987; Robertson et al., 1991; Robertson & Kandola, 1982; Rynes & Connerley, 1993; Schmitt et al., 1993; Schuler, 1993a; Smither et al., 1993; Stone-Romero et al., 2003). Die hohe Akzeptanz basiert darauf, dass diese beiden Verfahren einige wichtige Fairness-Kriterien von Gilliland (1993) erfüllen: Sie haben einen hohen Tätigkeitsbezug, die Bewerber haben die Möglichkeit zur Selbstdarstellung, die Durchführung ist standardisiert und damit vergleichbar und die Bewerber erhalten in der Regel ein spezifisches, leistungsbezogenes Feedback mit Entwicklungshinweisen.

Interviews sind valide Prädiktoren, wenn sie strukturiert durchgeführt werden (z. B. McDaniel, Whetzel, Schmidt & Maurer, 1994; Moscoso, 2000) und Bewerber stufen sie im Allgemeinen als sehr fair ein (z. B. Gilliland & Steiner, 1999; Kravitz et al., 1996), was sich auch in diesem Fall auf die Erfüllung einiger Fairness-Kriterien zurückführen lässt: Interviews bekommen eine hohe Augenscheinvalidität attestiert, bieten den Bewerbern die Möglichkeit zur Selbstdarstellung, ermöglichen es dem Interviewer, Informationen zum Auswahlverfahren und ein unmittelbares Feedback zu geben und stellen die prototypische Situation der Zweiweg-Kommunikation dar. Interessanterweise stufen Bewerber die Fairness von Interviews jedoch umso tiefer ein, je strukturierter – und somit valider – diese durchgeführt werden (z. B. Conway & Peneno, 1999; Kohn & Dipboye, 1998; Latham & Finnegan, 1993; Schuler, 1993b), was daran liegen könnte, dass die Strukturierung bei ihnen den Eindruck erweckt, dass man sie in der Möglichkeit zur Selbstdarstellung einschränkt. Die Erwartungshaltung der Bewerber, dass sie sich im Rahmen eines Selektionsverfahrens einem Interview zu stellen haben, ist so gross, dass dessen Ausbleiben zu einer tieferen Fairness-Einstufung des gesamten Selektionsverfahrens führt (Singer, 1992, 1993). Zudem sinkt die Akzeptanz bei den Bewerbern, wenn der Personalverantwortliche das Interview

anstelle im direkten Kontakt per Telefon durchführt (Chapman, Uggerslev & Webster, 2003).

Intelligenztests gelten als die validesten Prädiktoren beruflicher Leistung, besonders bei komplexen Tätigkeiten (z. B. Barrett & Depinet, 1991; Hartigan & Wigdor, 1989; Hunter & Hunter, 1984; Ree & Earles, 1992; Ree, Earles, & Teachout, 1994; Salgado, Anderson, Moscoso, Bertua & de Fruyt, 2003; Schmidt & Hunter, 1998a, b), werden aber von den Bewerbern nur mittelmässig geschätzt (Kravitz et al., 1996; Macan et al., 1994; Rynes & Connerley, 1993), wobei die Schwierigkeit des Tests (Kluger & Rothstein, 1993; Schmidt et al., 1977), das Abschneiden im Test (Chan, Schmitt, Sacco & DeShon, 1998; Kluger & Rothstein, 1993) und der Abstraktionsgrad der Aufgaben (Rosse, Miller, & Stecher, 1994; Smither et al., 1993) deren Akzeptanz-Beurteilung beeinflusst. So konnten Klingner und Schuler (2004) anhand eines Intelligenztests mit hohem Tätigkeitsbezug zeigen, dass die Studienteilnehmer diesen im Rahmen einer Bewerbungssituation deutlich besser akzeptieren als einen herkömmlichen Intelligenztest (56.4% vs. 20.5%). Auch akzeptieren die Bewerber Intelligenztests besser, wenn ihnen der Personalverantwortliche den Grund für den Einsatz des Testverfahrens erläutert (Horvath et al., 2000).

Persönlichkeitstests werden von den Bewerbern nicht sonderlich geschätzt (Harland, Rauzi & Biasotto, 1995; Kravitz et al., 1996; Rosse et al., 1994; Rynes & Connerley, 1993; Smither et al., 1993; Stone-Romero et al., 2003). Diese negative Einschätzung bleibt auch weit gehend erhalten, wenn der Selektionsverantwortliche den Bewerbern Erklärungen zur Testung abgibt (Harland et al., 1995; Rosse et al., 1996), wobei Informationen, welche aufzeigen warum ein Test durchgeführt wird, eine schwach positive Wirkung auf die Akzeptanzeinschätzung haben (Horvath et al., 2000). Jones (1991) konnte aufzeigen, dass Bewerber einen Persönlichkeitstest deutlich besser akzeptieren, wenn dessen Inhalte arbeitsbezogen sind, als wenn sie in Anlehnung an klinische Fragebogen formuliert sind. Dabei reicht es aber nicht aus, bei einem herkömmlichen Fragebogen bei jeder Aussage einfach noch den Zusatz „bei der Arbeit" hinzuzufügen, um eine höhere Einstufung des Tätigkeitsbezuges zu erzielen. Jedoch zeigte sich, dass der Hinweis, dass der Test eigens für diesen Selektionszweck entwickelt wurde und er die Arbeitsleistung gut vorhersagen kann, eine positive Auswirkung auf die Akzeptanz des Verfahrens hat (Holtz, Ployhart & Dominguez, 2005). Auch die Schaffung eines höheren Realitätsbezuges, indem zum Beispiel in einem Situational Judgment Test anstelle einer Beschreibung der Situation ein Videofilm eingesetzt wird, kann die Augenscheinvalidität erhöhen (positiv bei Chan & Schmitt, 1997; Richman-Hirsch, Olson-Buchanan & Drasgow, 2000; kein Unterschied bei Kanning et al., 2006).

Wie komplex der Prozess der Akzeptanzbeurteilung bei den Bewerbern ist, zeigen nachfolgende Studienergebnisse: So spielt es eine Rolle, mit welchen anderen Verfahren zusammen die Bewerber den Persönlichkeitstest zu absolvieren haben: Zum Beispiel wirkt sich die Kombination mit einem Interview *und* einem Intelligenztest akzeptanzsteigernd aus (Rosse et al., 1994). In der Studie von Ni und Hauenstein (1998) zeigte sich, dass der Grad des Eindringens in die Privatsphäre nur dann einen Effekt auf die Zufriedenheit mit dem Selektionsverfahren hatte, wenn die Bewerber gleichzeitig die Augenscheinvalidität als tief einstuften. Rafaeli (1999) zeigte in ihrem Simulationsexperiment zudem auf, dass eine sehr ausführliche Testung der Persönlichkeit (8 Stunden) zu einer besseren Gesamtbewertung des Selektionsprozesses führt, als eine kurze (2 Stunden). Jenkins und Griffith (2004) konnten belegen, dass die Bewerber einen massgeschneiderten Persönlichkeits-Fragebogen besser akzeptieren als einen Test, welcher breitgefasste Konstrukte misst. Nicht empfehlenswert ist es, den Bewerber darauf aufmerksam zu machen, dass einige Fragen als zu persönlich erscheinen mögen und dem Besorgnis Ausdruck zu geben, dass er sich bei der Bearbeitung des Fragebogens unwohl fühlen könnte. Dies führt dazu, dass der Test deutlich schlechter akzeptiert wird (Ambrose & Rosse, 2003).

Viele der oben aufgeführten Studien belegen, dass die Beurteilung der Akzeptanz von Testverfahren multikausal bedingt ist und nicht nur vom einzelnen Testverfahren per se abhängt. So zeigten zum Beispiel Elkins und Phillips (2000) auf, dass der Tätigkeitsbezug zusammen mit dem Selektionsentscheid einen Einfluss auf die Akzeptanz des eingesetzten Testverfahrens hat. Die Aussage, dass Bewerber ein Testverfahren generell besser akzeptieren als ein anderes, würde somit der Komplexität der Realität nicht vollständig gerecht (Truxillo et al., 2001): Bewerber können eine Simulation bezüglich Tätigkeitsbezug, einen Multiple-Choice-Test jedoch bezüglich Vergleichbarkeit der Durchführung fairer einstufen als andere Verfahren. Weiter gilt es zu beachten, dass abgewiesene Bewerber den Selektionsprozess und die dabei eingesetzten Testverfahren generell schlechter akzeptieren als aufgenommene (Chan et al., 1997; Chan, Schmitt, Sacco & DeShon, 1998; Gilliland, 1994; Noe & Steffy, 1987; Ployhart & Ryan, 1998; Ryan et al., 2000; Thorsteinson & Ryan, 1997; Truxillo & Bauer, 1999). Eine Möglichkeit, die Fairnesseinstufung abgewiesener Bewerber zu erhöhen, besteht darin, diesen vorgängig Informationen über den Selektionsprozess und Übungsmaterial zur Verfügung zu stellen (Burns et al., 2008).

Will nun ein Personalverantwortlicher die Akzeptanz seines Selektionsprozesses Erhöhen, genügt es also nicht, wenn er weiss, dass ein Test besser

akzeptiert wird als ein anderer, er muss auch Kenntnis davon haben, warum dies so ist und wie sich die Fairnesswahrnehmungen der einzelnen Selektionsverfahren gegenseitig beeinflussen (Madigan & Macan, 2005). Als die vier wichtigsten Faktoren, welche sich auf die Akzeptanz von Selektionsverfahren auswirken, nennen Anderson, Born und Cunningham-Snell (2001) den Tätigkeitsbezug, den Schutz der Privatsphäre, die Übereinstimmung mit den Erwartungen des Bewerbers bezüglich Verfahrens- oder Verteilungsgerechtigkeit und die Möglichkeit des persönlichen Kontaktes mit dem Personalverantwortlichen.

Kritisch ist anzufügen, dass viele der Studien zur Akzeptanz von Testverfahren nicht in realen Settings und mit Studenten durchgeführt wurden (Anderson, 2003). Zudem wurden selten Längsschnittstudien durchgeführt, um die Einstellung der Bewerber vor der Testung und nach dem Feedback zu kontrollieren. Auch wurde kaum – wie zum Beispiel von Borman, Hanson und Hedge (1997) gefordert – theoriegeleitet geforscht. Um bei der Messung der Akzeptanz diagnostischer Verfahren gültige Aussagen zu erhalten, schlägt Kersting (2008) fünf Postulate vor:

1. Eine Person kann die Akzeptanz eines Verfahrens nur einschätzen, wenn sie dieses auch bearbeitet hat (siehe dazu auch Marcus, 2003).
2. Zur Messung der Akzeptanz eines Verfahrens ist ein auf Theorien basierender, mehrdimensionaler, reliabler, valider Fragebogen einzusetzen.
3. Da psychologische Diagnostik in einer bestimmten Situation und einem spezifischen Kontext stattfindet, müssen diese Variablen bei der Messung der Akzeptanz eines Verfahrens berücksichtigt werden (z. B. Freiwilligkeit, Anonymität, Ziel und Bedeutsamkeit der Diagnose, erzieltes Ergebnis).
4. Als wichtige Moderatorvariablen der Akzeptanz sind auch die Eigenschaften der Person (z. B. demografische Variablen, Fähigkeit, Persönlichkeit, Interesse, Motivation), das erzielte Ergebnis und die Selbsteinschätzung der Leistung zu berücksichtigen.
5. Interessiert der Vergleich der Akzeptanz verschiedener Verfahren, so müssen diese in einem möglichst ähnlichen Setting durchgeführt und beurteilt werden.

Abschliessend gehe ich noch auf ein Problem ein, welches die Forschung zur Bewerberperspektive seit jeher beschäftigt: Cropanzano (1994; siehe auch Cropanzano & Konovsky, 1996) kam auf der Basis der dazumal vorliegenden Studienergebnissen zum Schluss, dass ein *„justice dilemma*" vorliegt, welches

dadurch entsteht, dass die Bewerber psychometrisch valide Testverfahren – zum Beispiel Intelligenz- und Integritätstests – tendenziell als unfairer einstufen als wenig valide Testverfahren, wie zum Beispiel ein unstrukturiertes Interview. Cropanzano und Wright (2003) führen dafür vier Gründe auf: Die Bewerber kennen die Bedeutung der Validität nicht, sie taxieren eine Verletzung ethisch-moralischer Richtlinien durch Testverfahren als gravierend, sie können durch valide Testverfahren erhobene, negative Informationen als selbstwertbedrohlich empfinden (Van den Bos, Bruins, Wilke & Dronkert, 1999) oder sie unterscheiden bei der Beurteilung der Auswirkungen eines Testverfahrens zwischen einem Mikro- (positive Konsequenzen für eine Person) und einem Makrolevel (negative Konsequenzen für die Gesellschaft auf Grund einer systematischen Benachteiligung einzelner Personengruppen). Wie ich in der nachfolgenden Tabelle 5.9 aufzeige, widerlegen die oben dargestellten Studien (wie auch andere Studien z. B. Kravitz et al., 1996; Rynes & Connerley, 1993) jedoch weit gehend dieses Validitätsdilemma, da es nur bei Integritätstests aufzutreten scheint.

Tabelle 5.9

Vergleich der Akzeptanzeinstufungen und der Validitäten von in der Personalselektion eingesetzten Testverfahren

Selektionsverfahren	Interview	Arbeitsprobe	Lebenslauf	Intelligenztests	biografischer Fragebogen	Referenzen und Zeugnisse	Persönlichkeitstests	Integritätstests	persönliche Kontakte	graphologisches Gutachten
Augenscheinvalidität	2	2	2	1	1	1	1	0	0	0
Möglichkeit zur Selbstdarstellung	2	2	2	2	1	1	1	1	0	0
Wahrung der Privatsphäre	2	2	2	1	1	1	1	1	1	0
Total	6	6	6	4	3	3	3	2	1	0
Mittelwerte aus den Studien	5.49	5.51	5.24	4.95	4.67	4.48	4.60	4.03	3.41	2.85
Validität (Schmidt & Hunter, 1998b; S. 22)	.38/.51	.54		.51	.35	.26	.31	.41		.02

Anmerkung. Für eine bessere Übersichtlichkeit habe ich die Akzeptanz-Werte vereinfacht dargestellt: Werte >5 als 2; Werte zwischen 4 und 5 als 1 und Werte <4 als 0.

Cropanzano und Wright (2003) nennen vier Vorschläge, um die Verfahrensgerechtigkeit bei Selektionsverfahren zu erhöhen:

- Ersetzen von validen, aber unfair beurteilten Testverfahren durch solche mit ähnlicher oder höherer Validität, welche Bewerber als fair beurteilen.
- Überarbeiten von als unfair eingestuften Testverfahren. Hier kann schon viel erreicht werden, wenn zu persönliche Fragen gestrichen werden und der Test einen deutlicheren Tätigkeitsbezug aufweist (Jones, 1991).
- Umfassende und einfühlsame Rückmeldung zum Entscheidungsprozess und der getroffenen Entscheidung. Zudem lässt sich die Meinung des Bewerbers über die eingesetzten Testverfahren verbessern, wenn der Personalverantwortliche diesen über die prädiktive Validität der eingesetzten Testverfahren informiert (Rynes & Connerley, 1993).
- Kombination verschiedener Selektionsverfahren. Wie weiter oben schon erwähnt, lassen sich zum Beispiel im Interview prädiktiv valide Fragen mit solchen mit positiveren Bewerberreaktionen mischen (Gilliland & Steiner, 1999).

Zudem schlagen Schleicher et al. (2006) vor, das Testmaterial dem Bewerberpool anzupassen, um so deren Möglichkeiten zur Selbstdarstellung zu optimieren und so schlussendlich einen als fairer wahrgenommenen Selektionsprozess zu erhalten. Dazu muss man die Charakteristiken des Bewerberpools – zum Beispiel den Bildungsgrad oder die Vorerfahrungen – studieren. Das Testmaterial ist dann so zu gestalten, dass es sich mit diesen Charakteristiken überlappt.

5.8 Die bewerberzentrierte Personalauswahl

Es ist bei der Personalauswahl in der betrieblichen Praxis wohl nur schwer möglich, alle der Gilliland-Regeln umzusetzen: Abgesehen davon, dass das Wissen darüber bei vielen Personalverantwortlichen wohl gar nicht aktiv präsent ist, fehlt das entsprechend geschulte Personal. Zudem wird eine konsequente Umsetzung der Fairness-Aspekte an den die Einführung verursachenden Kosten, den zusätzlichen zeitlichen Aufwand bei der Durchführung des Selektionsprozesses – zum Beispiel für ein umfassendes Feedback an den Bewerber – oder aus Ermangelung anhand dieser Kriterien entwickelten Testverfahren scheitern. Dabei würde sich eine bewerberzentrierte Personalauswahl (Boss, 2005) – Köchling (2000) spricht von Bewerberorientierung – mittelfristig für die Organisation auszahlen, da sich dieses Vorgehen auf vier Ebenen auswirkt:

- Es ist gewährleistet, dass die gesetzlichen Rahmenbedingungen eingehalten werden.
- Die Organisation kann sich in einem guten Licht und als interessanter Arbeitgeber präsentieren.
- Die Durchführungsverantwortlichen können sicher gehen, dass sie nach den Regeln der Kunst arbeiten.
- Der Bewerber empfindet sich als angenehm und fair behandelt.

Dass dieses Wissen noch nicht bis zu den Selektionsverantwortlichen durchgedrungen ist, hängt zu einem Teil auch damit zusammen, dass diese Thematik in der entsprechenden Fachliteratur – wenn überhaupt – nur am Rande behandelt wird. Hier sind Psychologen gefragt, welche die Forschungsergebnisse zusammentragen und in die Alltagssprache und die Praxisrealität transferieren. Dass dies nicht einfach ist, zeige ich an folgendem Beispiel: LaHuis et al. (2003) stellten fest, dass eine Erklärung, wozu der Einsatz eines Intelligenztests dient, zu einer höheren Fairnesswahrnehmung bei Bewerbern führt. Wie sich in der Studie von Harland et al. (1995) zeigte, trifft dies jedoch nicht für Persönlichkeits-Fragebogen zu, da die Bewerber diese anscheinend zu negativ wahrnehmen. Daraus eine Regel für die Personalverantwortlichen abzuleiten, welche besagt, wann zu welchem Testverfahren welche Erklärung abgegeben werden soll, wäre praxisfern und so nicht umsetzbar. Vielmehr geht es darum, Minimalstandards festzulegen, deren positive Auswirkungen auf den Bewerber und die Organisation mit wissenschaftlichen Studien belegt sind und welche sich problemlos im Selektionsalltag umsetzen lassen.

Im Kreise der Psychologinnen und Psychologen wird oft und mit Nachdruck betont, dass Betroffene zu Beteiligten gemacht werden müssen. Dies kann im Falle der Personalselektion beispielsweise so erreicht werden, indem die Aufgaben Testkandidaten zur Einschätzung der Akzeptanz, Verständlichkeit etc. vorgelegt werden oder dass die Personalfachkraft beim Feedbackgespräch ein echtes Interesse an der Sicht- und Erlebensweise des Bewerbers zeigt. Die Verantwortlichen in den Rekrutierungszentren der Schweizer Armee führen vor der Entlassung aus der Rekrutierung eine Meinungsumfrage durch, bei welcher jeder Stellungspflichtige auf einem Fragebogen seine (Un-)Zufriedenheit mit den Abläufen, der Verpflegung und der Unterkunft oder der medizinischen und psychologischen Untersuchungen mitteilen kann. Die aggregierten Antworten dienen im Rahmen der Qualitätskontrolle dazu, gezielte Verbesserungen einzuführen. Aussagekräftiger wäre eine Erhebung der Fairness des Selektionsprozesses unmittelbar nach der Testung und dann ein zweites Mal nach dem Feedbackgespräch, da das Abschneiden im Selektionsverfahren einen Einfluss auf dessen Beurteilung hat (Schleicher et al., 2006).

Eines darf aber bei allen schönen Worten und Vorsätzen nicht vergessen werden: Eine echt partnerschaftliche Entscheidungsfindung kann es in einer Selektionssituation oft nicht geben. In den meisten Fällen besteht ein Machtgefälle zwischen Diagnostiker und Bewerber und am Schluss müssen die Personalverantwortlichen auch Entscheide gegen die Interessen eines Bewerbers treffen, da Schuler und Stehles (1983) Maxime der Selbstselektion durch Information vor allem bei Arbeitsplatzknappheit oder prestigeträchtigen Positionen wohl kaum greifen wird.

Abschliessend führe ich noch eine mir persönlich mitgeteilte Anekdote auf, welche zeigt, welche Auswirkungen der unreflektierte Einsatz von Verfahren in einem Selektionsprozess haben kann: Einer meiner ehemaligen Studenten bewarb sich auf eine Stelle als Psychologe in der Assessment-Abteilung einer Bank. Im Verlaufe des Selektionsverfahrens hatte er auch ein Rollenspiel zu absolvieren, in welchem er einen Verkäufer zu spielen hatte, der ein bestimmtes Produkt einem schwierigen Kunden anzupreisen hatte. Schon bei der Instruktion zu diesem Rollenspiel überlegte er sich, ob er aus der Selektion aussteigen soll, weil er den Fall als so unpassend für seine spätere Tätigkeit bei dieser Bank erlebte, „spielte" dann aber doch mit, weil er noch ein Feedback zu seinen Leistungen haben wollte. Die ganze Situation erschien ihm jedoch derart unprofessionell, dass er sich entschied nicht in diesem Team arbeiten zu wollen, obwohl ihn die Aufgabe interessierte und er dafür bestens qualifiziert gewesen wäre.

5.9 Literaturverzeichnis

Adams, J. S. (1963). Toward an understanding of inequity. *Journal of Abnormal and Social Psychology, 67*, 422–436.

Ambrose, M. L., & Cropanzano, R. (2003). A longitudinal analysis of organizational fairness: An examination of reactions to tenure and promotion decisions. *Journal of Applied Psychology, 88*, 266–275.

Ambrose, M. L., & Rosse, J. G. (2003). Procedural justice and personality testing: An examination of concern and typicality. *Group & Organization Management, 28*, 502–526.

Amrine, M. (1965). The 1965 congressional inquiry into testing: A commentary. *American Psychologist, 20*, 859–870.

Anderson, N. (2003). Applicant and recruiter reactions to new technology in selection: A critical review and agenda for future research. *International Journal of Selection and Assessment, 11*, 121–136.

Anderson, N. (2004). Editorial – The dark side of the moon: Applicant perspectives, negative psychological effects (NPEs), and candidate decision making in selection. *International Journal of Selection and Assessment, 12*, 1–8.

Anderson, N., Born, M., & Cunningham-Snell, N. (2001). Recruitment and selection: Applicant perspectives and outcomes. In: N. Anderson, D. S. Ones, H. K. Senangil, & C. Viswesvaran (Eds.), *Handbook of industrial, work and organizational psychology* (Vol. 1., pp. 200–218). London: Sage.

Anderson, N., & Witvliet, C. (2008). Fairness reactions to personnel selections methods: An international comparison between the Netherlands, the United States, France, Spain, Portugal, and Singapore. *International Journal of Selection and Assessment, 16*, 1–13.

Anseel, F., & Lievens, F. (2009). The mediating role of feedback acceptance in the relationship between feedback and attitudinal and performance outcomes. *International Journal of Selection and Assessment, 17*, 362–376.

Arvey, R. D., & Sackett, P. R. (1993). Fairness in selection: Current developments and perspectives. In N. Schmitt & W. Borman (Eds.), *Personnel selection in organizations* (pp. 171–202). San Francisco; CA: Jossey-Bass.

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. *Personnel Psychology, 43*, 695–716.

Bangerter, A., König, C. J., Blatti, S., & Salvisberg, A. (2009). How widespread is graphology in personnel selection practice? A case study of a job market myth. *International Journal of Selection and Assessment, 17*, 219–230.

Barrett, G. V., & Depinet, R. L. (1991). A reconsideration of testing for competence rather than for intelligence. *American Psychologist, 46*, 1012–1024.

Bauer, T. N., Maertz, C. P., Dolen, M. R., & Campion, M. A. (1998). Longitudinal assessment of applicant reactions to employment testing and test outcome feedback. *Journal of Applied Psychology, 83*, 892–903.

Bauer, T. N., Truxillo, D. M., & Paronto, M. E. (2003). The measurement of applicant reactions to selection-related events and outcomes. In J. C. Thomas (Series Ed.) & M. Hersen (Vol. Ed.), *Comprehensive handbook of psychological assessment: Vol. 4. Industrial and organizational assessment* (pp. 482–506). New York, NY: Wiley.

Bauer, T. N., Truxillo, D. M., Sanchez, R. J., Craig, J. M., Ferrara, Ph., & Campion, M. A. (2001). Applicant reactions to selection: Development of the selection procedural justice scale (SPJS). *Personnel Psychology, 54,* 387–419.

Bell, B. S., Ryan, A. M., & Wiechmann, D. (2004). Justice expectations and applicant perceptions. *International Journal of Selection and Assessment, 12*, 24–38.

Bell, B. S., Wiechmann, D., & Ryan, A. M. (2006). Consequences of organizational justice expectations in a selection system. *Journal of Applied Psychology, 91*, 455–466.

Bertolino, M., & Steiner, D. D. (2007). Fairness reactions to selection methods: An Italian study. *International Journal of Selection and Assessment, 15,* 197–205.

Bies, R. J. (1993). Privacy and procedural justice in organizations. *Social Justice Research, 6*, 69–86.

Bies, R. J. (2005). Are procedural justice and interactional justice conceptually distinct? In J. Greenberg & J. A. Colquitt (Eds.), *Handbook of organizational justice* (pp. 85–112). Mahwah, NJ: Erlbaum.

Bies, R. J., & Moag, J. S. (1986). Interactional justice: Communication criteria of fairness. *Research on Negotiation in Organizations, 1*, 43–55.

Bies, R. J., & Shapiro, D. L. (1987). Interactional fairness judgments: The influence of causal accounts. *Social Justice Research, 1*, 199–218.

Bies, R. J., & Shapiro, D. L. (1988). Voice and justification: Their influence on procedural judgments. *Academy of Management Journal, 31*, 676–685.

Bies, R. J., Shapiro, D. L., & Cummings, L. L. (1988). Causal accounts and managing organizational conflict: Is it enough to say it's not my fault? *Communication Research, 15*, 381–399.

Bies, R. J., & Tyler, T. R. (1993). The ''litigation mentality'' in organizations: A test of alternative psychological explanations. *Organization Science, 4*, 352–366.

Borchers, M. (1986). *Zur Akzeptanz von Persönlichkeitsverfahren in der Personalberatung.* Unveröff. Diplomarbeit, Universität Köln.

Borman, W. C., Hanson, M. A., & Hedge, J. W. (1997). Personnel selection. *Annual Review of Psychology, 48*, 299–337.

Boss, P. (2005). Assessment in der Arbeitswelt – Kriterien für eine bewerberzentrierte Personalauswahl. In M. Rehbinder (Hrsg.), *Psychologische Aspekte im Recht der Personalführung* (S. 21–45). Bern: Stämpfli.

Boss, P. & Baumann, R. (2003). Psychologische Testverfahren beim Rekrutierungsprozess der Armee. *HR-Today, 7–8*, 22–23.

Boudreau, J. W., & Rynes, S. L. (1985). The role of recruitment in staffing utility analysis. *Journal of Applied Psychology, 70*, 354–366.

Bourgeois, R. P., Leim, M. A., Slivinski, L. W., & Grant, K. W. (1975). Evaluation of an assessment center in terms of acceptability. *Canadian Personnel and Industrial Relations Journal, 22*, 17–20.

Brett, J. F., & Atwater, L. E. (2001). 360° Feedback: Accuracy, reactions, and perceptions of usefulness. *Journal of Applied Psychology, 86*, 930–942.

Bretz, R. D., & Judge, T. A. (1998). Realistic job previews: A test of the adverse self-selection hypothesis. *Journal of Applied Psychology, 83*, 330–337.

Brockner, J., Tyler, T. R., & Cooper-Schneider, R. (1992). The influence of prior commitment to an institution on reactions to perceived unfairness: The higher they are, the harder they fall. *Administrative Science Quarterly, 37*, 317–348.

Burns, G. N., Siers, B. P., & Christiansen, N. D. (2008). Effects of providing pretest information and preparation materials on applicant reactions to selection procedures. *International Journal of Selection and Assessment, 16*, 73–77.

Byrne, Z. S., & Cropanzano, R. (2001). The history of organizational justice: The founders speak. In R. Cropanzano (Ed.), *Justice in the workplace. From theory to practice* (Vol. 2, pp. 3–26). Mahwah, NJ: Erlbaum.

Cascio, W. F., & Aguinis, H. (2005). *Applied psychology in human resource management* (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Chan, D. (1997). Racial subgroup differences in predictive validity perceptions on personality and cognitive ability tests. *Journal of Applied Psychology, 82*, 311–320.

Chan, D., & Schmitt, N. (1997). Video-based versus paper-and-pencil method of assessment in situational judgment tests: Subgroup differences in test performance and face validity perceptions. *Journal of Applied Psychology, 82*, 143–159.

Chan, D., & Schmitt, N. (2004). An agenda for future research on applicant reactions to selection procedures: A construct-oriented approach. *International Journal of Selection and Assessment, 12*, 9–23.

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology, 82*, 300–310.

Chan, D., Schmitt, N., Jennings, D., Clause, C. S., & Delbridge, K. (1998). Applicant perceptions of test fairness: Integrating justice and self-serving bias perspectives. *International Journal of Selection and Assessment, 6*, 232–239.

Chan, D., Schmitt, N., Sacco, J. M., & DeShon, R. P. (1998). Understanding pretest and posttest reactions to cognitive ability and personality tests. *Journal of Applied Psychology, 83*, 471–485.

Chapman, D. S., Uggerslev, K. L., & Webster, J. (2003). Applicant reactions to face-to-face and technology-mediated interviews: A field investigation. *Journal of Applied Psychology, 88*, 944–953.

Civil Rights Act, 42 U.S.C. § 2000e et seq. (1964).

Clause, C. S., Delbridge, K., Schmitt, N., Chan, D., & Jennings, D. (2001). Test preparation activities and employment test performance. *Human Performance, 14*, 149–167.

Cohen-Charash, Y., & Spector, P. E. (2001). The role of justice in organizations: A meta-analysis. *Organizational Behavior and Human Decision Processes, 86*, 278–321.

Colquitt, J. A. (2001). On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology, 86*, 386–400.

Colquitt, J. A., & Chertkoff, J. M. (2002). Explaining injustice: The interactive effect of explanation and outcome on fairness perceptions and task motivation. *Journal of Management, 28*, 591–610.

Colquitt, J. A., Colon, D. E., Wesson, M. J., Porter, C. O., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology, 86*, 425–445.

Colquitt, J. A., Greenberg, J., & Zapata-Phelan, C. P. (2005). What is organizational justice? A historical overview. In J. Greenberg & J. A. Colquitt (Eds.), *Handbook of organizational justice* (pp. 3–56). Mahwah, NJ: Erlbaum.

Colquitt, J. A., & Shaw, J. C. (2005).How should organizational justice be measured? In J. Greenberg & J. A. Colquitt (Eds.), *Handbook of organizational justice* (pp. 113–152). Mahwah, NJ: Erlbaum.

Converse, P. D., Oswald, F. L., Imus A., Hedricks, C., Roy, R., & Butera, H. (2008). Comparing personality test formats and warnings: Effects on criterion-related validity and test-taker reactions. *International Journal of Selection and Assessment, 16*, 155–169.

Conway, J. M., & Peneno, G. M. (1999). Comparing structured interview question types: Construct validity and applicant reactions. *Journal of Business and Psychology, 13*, 485–505.

Cook, M. (2004). *Personnel selection. Adding value through people* (4th ed.). Chichester: Wiley.

Crant, J. M., & Bateman, T. S. (1990). An experimental test of the impact of drug-testing programs on potential job applicants' attitudes and intentions. *Journal of Applied Psychology, 75*, 127–131.

Cropanzano, R. (1994). The justice dilemma in employee selection: Some reflections on the trade-offs between fairness and validity. *The Industrial–Organizational Psychologist, 31*, 90–93.

Cropanzano, R., Bowen, D. E., & Gilliland, S. W. (2007). The management of organizational justice. *Academy of Management Perspectives, 21*, 34–48.

Cropanzano, R., & Konovsky, M. A. (1996). Resolving the justice dilemma by improving the outcomes: The case of employee drug screening. *Journal of Business and Psychology, 11*, 239–263.

Cropanzano, R., & Wright, T. A. (2003). Procedural justice and organizational staffing: A tale of two paradigms. *Human Resource Management Review, 13*, 7–39.

Daly, J. P., & Geyer, P. D. (1994). The role of fairness in implementing large-scale change: Employee evaluations of process and outcome in seven facility relocations. *Journal of Organizational Behavior, 15*, 623–638.

Derous, E., Born, M. Ph., & De Witte, K. (2004). How applicants want and expect to be treated: Applicants' selection treatment beliefs and the development of the Social Process Questionnaire on Selection. *International Journal of Selection and Assessment, 12,* 99–119.

Derous, E., & De Witte, K. (2001). Looking at selection from a social process perspective: Towards a social process model on personnel selection. *European Journal of Work and Organizational Psychology, 10*, 319–342.

Derous, E., De Witte, K., & Stroobants, R. (2003). Testing the social process model on selection through expert analysis. *Journal of Occupational and Organizational Psychology, 76*, 179–199.

Derous, E., & Schreurs, B. (2009). Modeling the structure of applicant reactions: An empirical study within the Belgian military. *Military Psychology, 21*, 40–61.

Deutsch, M. (1975). Equity, equality, and need: What determines which value will be used as the basis of distributive justice? *Journal of Social Issues, 31*, 137–149.

DIN (2002). *DIN 33430: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen*. Berlin: Beuth.

Dineen, B. R., Noe, R. A., & Wang, C. (2004). Perceived fairness of web-based applicant screening procedures: Weighing the rules of justice and the

role of individual differences. *Human Resource Management, 43*, 127–145.

Dipboye, R. L., & de Pontbriand, R. (1981). Correlates of employee reactions to performance appraisals and appraisal systems. *Journal of Applied Psychology, 66*, 248–251.

Donovan, M. A., Drasgow, F., & Munson, L. J. (1998). The Perceptions of Fair Interpersonal Treatment scale: Development and validation of a measure of interpersonal treatment in the workplace. *Journal of Applied Psychology, 83*, 683–692.

Elkins, T. J., & Phillips, J. S. (2000). Job context, selection decision outcome, and the perceived fairness of selection tests: Biodata as an illustrative case. *Journal of Applied Psychology, 85*, 479–484.

Equal Employment Opportunity Act of 1972, Pub. L. No. 92-261, 86 Stat. 103 (1972).

Fiske, D. W. (1967). The subject reacts to tests. *American Psychologist, 22*, 287–296.

Folger, R., & Greenberg, J. (1985). Procedural justice: An interpretive analysis of personnel systems. In K. M. Rowland & G. R. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 3, pp. 141–183). Greenwich, CT: JAI.

Folger, R., & Konovsky, M. A. (1989). Effects of procedural and distributive justice on reactions to pay raise decisions. *Academy of Management Journal, 32*, 115–130.

Fruhner, R., Schuler, H., Funke, U. & Moser, K. (1991). Einige Determinanten der Bewertung von Personalauswahlverfahren. *Zeitschrift für Arbeits- und Organisationspsychologie, 35,* 170–178.

Gatewood, R. D., Feild, H. S., & Barrick, M. (2008). *Human resource selection* (6th ed.). Mason, OH: Thomson Higher Education.

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *The Academy of Management Review, 18*, 694–734.

Gilliland, S. W. (1994). Effects of procedural and distributive justice on reactions to a selection system. *Journal of Applied Psychology, 79*, 691–701.

Gilliland, S. W. (1995). Fairness from the applicant's perspective: Reactions to employee selection procedures. *International Journal of Selection and Assessment, 3*, 11–18.

Gilliland, S. W., & Beckstein, B. A. (1996). Procedural and distributive justice in the editorial review process. *Personnel Psychology, 49*, 669–691.

Gilliland, S. W., Groth, M., Baker, R. C., Dew, A. F., Polly, L. M., & Langdon, J. C. (2001). Improving applicants' reactions to rejection letters: An application of fairness theory. *Personnel Psychology, 54*, 669–703.

Gilliland, S. W., & Hale, J. M. (2005). How can justice be used to improve employee selection practices? In J. Greenberg & J. A. Colquitt (Eds.), *Handbook of organizational justice* (pp. 411–438). Mahwah, NJ: Erlbaum.

Gilliland, S. W., & Honig, H. (1994, April). *Development of the selection fairness survey.* Paper presented at the 9th Annual Conference of the Society for Industrial and Organizational Psychology, Nashville, TN.

Gilliland, S. W., & Steiner, D. D. (1999). Applicant reactions. In R. W. Eder & M. M. Harris (Eds.), *The employment interview handbook* (pp. 69–86). Thousand Oaks, CA: Sage.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory. Strategies for qualitative research*. Chicago, IL: Aldine.

Graczyk, A. J. (2005). An examination of the influence of formal characteristics of paper-and-pencil selection test procedures on outcomes: A policy-capturing approach. *Dissertation Abstracts International, 66 (3-B) 2005,* 1770.

Greenberg, J. (1986). Determinants of perceived fairness of performance evaluations. *Journal of Applied Psychology, 71*, 340–342.

Greenberg, J. (1987). A taxonomy of organizational justice theories. *Academy of Management Review, 12*, 9–22.

Greenberg, J. (1990). Organizational justice: Yesterday, today, and tomorrow. *Journal of Management, 16*, 399–432.

Greenberg, J. (1993a). The social side of fairness: Interpersonal and informational classes of organizational justice. In R. Cropanzano (Ed.), *Justice in the workplace: Approaching fairness in human resource management* (pp. 79–103). Hillsdale, NJ: Erlbaum.

Greenberg, J. (1993b). Stealing in the name of justice: Informational and interpersonal moderators of theft reactions to underpayment inequity. *Organizational Behavior and Human Decision Processes, 54*, 81–103.

Grubitzsch, S. (1978). Sozialökonomische Grundlagen des Testens und Messens. In S. Grubitzsch & G. Rexilius (Hrsg.), *Testtheorie – Testpraxis. Voraussetzungen, Verfahren, Formen und Anwendungsmöglichkeiten psychologischer Tests im kritischen Überblick* (S. 40–51). Reinbek: Rowohlt.

Grubitzsch, S. & Rexilius, G. (Hrsg.). (1978). *Testtheorie – Testpraxis. Voraussetzungen, Verfahren, Formen und Anwendungsmöglichkeiten psychologischer Tests im kritischen Überblick*. Reinbek: Rowohlt.

Guion, R. M. (1967). Personnel selection. *Annual Review of Psychology, 18*, 191–216.

Häcker, H., Leutner, D. & Amelang, M. (Hrsg.). (1998). *Standards für pädagogisches und psychologisches Testen.* Bern: Huber.

Harburger, W. (1992). Soziale Validität im individuellen Erleben von Assessment-Center-Probanden. *Zeitschrift für Arbeits- und Organisationspsychologie, 36,* 147–151.

Harland, L. K., Rauzi, T., & Biasotto, M. M. (1995). Perceived fairness of personality tests and the impact of explanations for their use. *Employee Responsibilities and Rights Journal, 8*, 183–192.

Harn, T. J., & Thornton, G. C., III. (1985). Recruiter counselling behaviours and applicant impressions. *Journal of Occupational Psychology, 58*, 57–65.

Harold, C. M., & Ployhart, R. E. (2008). What do applicants want? Examining changes in attribute judgments over time. *Journal of Occupational and Organizational Psychology, 81*, 191–218.

Hartigan, J. A., & Wigdor, A. K. (Eds.). (1989). *Fairness in employment testing: Validity generalization, minority issues and the General Aptitude Test Battery.* Washington, DC: National Academy Press.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*, 639–683.

Hehlen, H. (1978). *Selektion. Aufsteigen, Absteigen, Beharren: Bildung als Herrschaftsmittel des Menschen über den Menschen am Beispiel einer Mittelschule.* Zürich: Verlagsgenossenschaft.

Heneman, H. G., III. (1985). Pay satisfaction. In K. Rowland & G. Ferris (Eds.), *Research in personnel and human resources management* (Vol. 3, pp. 115–139). Greenwich, CT: JAI.

Herberger, J. (1984). *Partizipation und Eignungsdiagnostik. Ein empirischer Beitrag zum Konzept der sozialen Validität.* Unveröff. Diplomarbeit, Universität Erlangen-Nürnberg.

Herriot, P. (1989). Selection as a social process. In M. Smith & I. Robertson (Eds.), *Advances in selection and assessment* (pp. 171–187). New York, NY: Wiley.

Holtz, B. C., Ployhart, R. E., & Dominguez, A. (2005). Testing the rules of justice: The effects of frame-of-reference and pre-test validity information on personality test responses and test perceptions. *International Journal of Selection and Assessment, 13*, 75–86.

Homans, G. S. (1961). *Social behavior: Its elementary forms*. New York, NY: Harcourt, Brace & World.

Hornke, L. F. & Winterfeld, U. (Hrsg.). (2004). *Eignungsbeurteilungen auf dem Prüfstand: DIN 33430 zur Qualitätssicherung.* Heidelberg: Spektrum Akademischer Verlag.

Horvath, M., Ryan, A. M., & Stierwalt, S. L. (2000). The influence of explanations for selection test use, outcome favorability, and self-efficacy on test-taker perceptions. *Organizational Behavior and Human Decision Processes, 83*, 310–330.

Hough, L. M., Oswald, F. L., & Ployhart, R. E. (2001). Determinants, detection and amelioration of adverse impact in personnel selection procedures: Issues, evidence and lessons learned. *International Journal of Selection and Assessment, 9*, 152–194.

Hülsheger, U. R., & Anderson, N. (2009). Applicant perspectives in selection: Going beyond preference reactions. *International Journal of Selection and Assessment, 17*, 335–345.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–98.

Iles, P. A., & Robertson, I. T. (1989). The impact of personnel selection procedures on candidates. In P. Herriot (Ed.), A*ssessment and selection in organizations: Methods and practices for recruitment and appraisal* (pp. 257–271). Chichester, UK: Wiley.

Iles, P. A., & Robertson, I. T. (1997). The impact of personnel selection procedures on candidates. In N. Anderson & P. Herriot (Eds.), *International handbook of selection and assessment. Assessment and selection in organizations, methods and practice for recruitment and appraisal* (Vol. 2, pp. 543–566). Chichester, UK: Wiley.

Ilgen, D. R., & Davis, C. A. (2000). Bearing bad news: Reactions to negative performance feedback. *Applied Psychology: An International Review, 49*, 550–565.

Ispas, D., Ilie, A., Iliescu, D., Johnson, R. E., & Harris, M. M. (2010). Fairness reactions to selection methods: A Romanian study. *International Journal of Selection and Assessment, 18*, 102–110.

Jäger, A. O. (1961). Personalauslese. In P. Lersch, F. Sander, H. Thomae (Hrsg. Serie) & A. Mayer, B. Herwig (Hrsg. Band), *Handbuch der Psychologie: Band 9. Betriebspsychologie* (S. 569–613). Göttingen: Verlag für Psychologie, Dr. C. J. Hogrefe.

Jenkins, M., & Griffith, R. (2004). Using personality constructs to predict performance: Narrow or broad bandwidth. *Journal of Business and Psychology, 19,* 255–269.

Jones, J. W. (1991). Assessing privacy invasiveness of psychological test items: Job relevant versus clinical measures of integrity. *Journal of Business and Psychology, 5*, 531–535.

Judge, T. A., & Colquitt, J. A. (2004). Organizational justice and stress: The mediating role of work-family conflict. *Journal of Applied Psychology, 89*, 395–404.

Kanning, U. P. (2004). *Standards der Personaldiagnostik.* Göttingen: Hogrefe.

Kanning, U. P., Grewe, K., Hollenberg, S., & Hadouch, M. (2006). From the subjects' point of view: Reactions to different types of situational judgment tests. *European Journal of Psychological Assessment, 22*, 168–176.

Katz, M. D., & Kahn, R. L. (1978). *The social psychology of organizations* (2nd ed). New York, NY: Wiley.

Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. *Annual Review of Psychology, 31*, 457–501.

Kersting, M. (1998). Differenzielle Aspekte der sozialen Akzeptanz von Intelligenztests und Problemlöseszenarien als Personalauswahlverfahren. *Zeitschrift für Arbeits- und Organisationspsychologie, 42*, 61–75.

Kersting, M. (Juni 2006). *Akzeptanz in der psychologischen Diagnostik.* Unterlagen zum Vortrag am Psychologischen Institut der Universität Zürich, Fachrichtung Persönlichkeitspsychologie und Diagnostik, Zürich.

Kersting, M. (2008). Zur Akzeptanz von Intelligenz- und Leistungstests. *Report Psychologie, 33*, 420–433.

Kersting, M., Althoff, K. & Jäger, A. O. (2008). *WIT-2. Der Wilde-Intelligenztest. Verfahrenshinweise.* Göttingen: Hogrefe.

Kirkpatrick, J. J., Ewen, R. B., Barrett, R. S., & Katzell, R. A. (1968). *Testing and fair employment. Fairness and validity of personnel tests for different ethnic groups.* New York, NY: New York University Press.

Klingner, Y., & Schuler, H. (2004). Improving participants' evaluations while maintaining validity by a work sample-intelligence test hybrid. *International Journal of Selection and Assessment, 12*, 120–134.

Kluger, A. N., & Rothstein, H. R. (1993). The influence of selection test type on applicant reactions to employment testing. *Journal of Business and Psychology, 8*, 3–25.

Köchling, A. C. (2000). *Bewerberorientierte Personalauswahl. Ein effektives Instrument des Personalmarketings.* Bern: Lang.

Kohn, L. S., & Dipboye, R. L. (1998). The effects of interview structure on recruiting outcomes. *Journal of Applied Social Psychology, 28*, 821–843.

Konovsky, M. A., & Cropanzano, R. (1991). The perceived fairness of employee drug testing as a predictor of employee attitudes and job performance. *Journal of Applied Psychology, 76*, 698–707.

Korman, A. K. (1970). Toward a hypothesis of work behavior. *Journal of Applied Psychology, 54*, 31–41.

Kravitz, D. A., Stinson, V., & Chavez, T. L. (1996). Evaluations of tests used for making selection and promotion decisions. *International Journal of Selection and Assessment, 4*, 24–34.

LaHuis, D. M., Perreault, N. E., & Ferguson, M. W. (2003). The effect of legitimizing explanations on applicants' perception of selection assessment fairness. *Journal of Applied Social Psychology, 33,* 2198–2215.

Latham, G. P., & Finnegan, B. J. (1993). Perceived practicality of unstructured, patterned, and situational interviews. In H. Schuler, J. L. Farr, & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 41–55). Hillsdale, NJ: Erlbaum.

Leventhal, G. S. (1976). The distribution of rewards and resources in groups and organizations. In L. Berkowitz & W. Walster (Ed.), *Advances in experimental social psychology* (Vol. 9, pp. 91–131). New York, NY: Academic Press.

Leventhal, G. S. (1980). What should be done with equity theory? In K. J. Gergen, M. S. Greenberg, & R. H. Willis (Eds.), *Social exchange: Advances in theory and research* (pp. 27–55). New York, NY: Plenum.

Leventhal, G. S., Karuza, J. & Fry, W. R. (1980). Es geht nicht nur um Fairness: Eine Theorie der Verteilungspräferenzen. In G. Mikula (Hrsg.), *Gerechtigkeit und soziale Interaktion: Experimentelle und theoretische Beiträge aus der psychologischen Forschung* (S. 185–250). Bern: Huber.

Liden, R. C., & Parsons, C. K. (1986). A field study of job applicant interview perceptions, alternative opportunities, and demographic characteristics. *Personnel Psychology, 39*, 109–122.

Lievens, F., De Corte, W., & Brysse, K. (2003). Applicant perceptions of selection procedures: The role of selection information, belief in tests, and comparative anxiety. *International Journal of Selection and Assessment, 11*, 67–77.

Lind, E. A., & Tyler, T. R. (1988). *The social psychology of procedural justice.* New York, NY: Plenum Press.

Lounsbury, J. W., Bobrow, W., & Jensen, J. B. (1989). Attitudes toward employment testing: Scale development, correlates, and "known-group" validation. *Professional Psychology: Research and Practice, 20*, 340–349.

Lülsdorf, C. (1986). Einstellung ehemaliger Teilnehmer zur Assessment-Center-Methode und Determinanten dieser Einstellung. Unveröff. Diplomarbeit, Universität Bonn.

Macan, T. H., Avedon, M. J., Paese, M., & Smith, D. E. (1994). The effects of applicants' reactions to cognitive ability tests and an assessment center. *Personnel Psychology, 47*, 715–738.

Madigan, J., & Macan, T. H. (2005). Improving applicant reactions by altering test administration. *Applied H.R.M. Research, 10*, 73–87.

Maier, G. W., Streicher, B., Jonas, E. & Woschée, R. (2007). Gerechtigkeitsschätzungen in Organisationen: Die Validität einer deutschsprachigen Fassung des Fragebogens von Colquitt (2001). *Diagnostica, 53*, 97–108.

Marcus, B. (2003). Attitudes towards personnel selection methods: A partial replication and extension in a German sample. *Applied Psychology: An International Review, 52,* 515–532.

McCarthy, J. M., & Goffin, R. D. (2003). Is the Test Attitude Survey psychometrically sound? *Educational and Psychological Measurement, 63*, 446–464.

McDaniel, M. A., Whetzel, D. L., Schmidt, F. L., & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology, 79*, 599–616.

McEnrue, M. P. (1989). The perceived fairness of managerial promotion practices. *Human Relations, 42*, 815–827.

McEvoy, G. M., & Cascio, W. C. (1985). Strategies for reducing employee turnover: A meta-analysis. *Journal of Applied Psychology, 70*, 342–353.

Meglino, B. M., Ravlin, E. C., & DeNisi, A. S. (2000). A meta-analytic examination of realistic job preview effectiveness: A test of three counterintuitive propositions. *Human Resource Management Review, 10*, 407–434.

Mehlman, R. C., & Snyder, C. R. (1985). Excuse theory: A test of the self-protective role of attributions. *Journal of Personality and Social Psychology, 49*, 994–1001.

Moscoso, S. (2000). Selection interview: A review of validity evidence, adverse impact and applicant reactions. *International Journal of Selection and Assessment, 8*, 237–247.

Moscoso, S., & Salgado, J. F. (2004). Fairness reactions to personnel selection techniques in Spain and Portugal. *International Journal of Selection and Assessment, 12*, 187–196.

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational and Psychological Measurement, 7*, 191–205.

Nevo, B. (1993). The practical and theoretical value of Examinee Feedback Questionnaires (EFeQ). In B. Nevo & R. S. Jäger (Eds.), *Educational and psychological testing: The test taker's outlook* (pp. 85–113). Bern: Hogrefe & Huber.

Nevo, B. (1995). Examinee Feedback Questionnaire: Reliability and Validity Measures. *Educational and Psychological Measurement, 55*, 499–504.

Nevo, B., & Jäger, R. S. (Eds.). (1993). *Educational and psychological testing: The test taker's outlook.* Bern: Hogrefe & Huber.

Nevo, B., & Sfez, J. (1985). Examinees' Feedback Questionnaires. *Assessment and Evaluation in Higher Education, 10*, 235–243.

Ni, Y., & Hauenstein, N. M. A. (1998). Applicant reactions to personality tests: Effects of item invasiveness and face validity. *Journal of Business and Psychology, 12*, 391–406.

Nikolaou, I., & Judge, T. A. (2007). Fairness reactions to personnel selection techniques in Greece: The role of core self-evaluations. *International Journal of Selection and Assessment, 15*, 206–219.

Noe, R. A., & Steffy, B. D. (1987). The influence of individual characteristics and assessment center evaluation on career exploration behavior and job involvement. *Journal of Vocational Behavior, 30*, 187–202.

Noon, A. L. (2006). *Job applicants' testing and organizational perceptions: The effects of test information and attitude strength.* Unpublished doctoral dissertation, University of Nebraska – Lincoln.

Nowakowski, J. M., & Conlon, D. E. (2005). Organizational justice: Looking back, looking forward. *The International Journal of Conflict Management, 16*, 4–29.

Packard, V. (1966). *Die wehrlose Gesellschaft.* München: Droemer Knaur.

Paczensky, S. von (1974). *Der Testknacker. Wie man Karriere-Tests erfolgreich besteht.* München: Bertelsmann.

Pawlik, K. (Hrsg.). (1976). *Diagnose der Diagnostik. Beiträge zur Diskussion der psychologischen Diagnostik in der Verhaltensmodifikation.* Stuttgart: Klett.

Phillips, J. M. (1998). Effects of realistic job previews on multiple organizational outcomes: A meta-analysis. *Academy of Management Journal, 41*, 673–690.

Phillips, J. M., & Gully, S. M. (2002). Fairness reactions to personnel selection techniques in Singapore and the United States. *International Journal of Human Resource Management, 13*, 1186–1205.

Ployhart, R. E., Holcombe Ehrhart, K., & Hayes, S. C. (2005). Using attributions to understand the effects of explanations on applicant reactions: Are reactions consistent with the covariation principle? *Journal of Applied Social Psychology, 35*, 259–296.

Ployhart, R. E., & Holtz, B. C. (2008). The diversity-validity dilemma: Strategies for reducing racioethnic and sex subgroup differences and adverse impact in selection. *Personnel Psychology, 61*, 153–172.

Ployhart, R. E., & Ryan, A. M. (1997). Toward an explanation of applicant reactions: An examination of organizational justice and attribution frameworks. *Organizational Behavior and Human Decision Processes, 72*, 308–335.

Ployhart, R. E., & Ryan, A. M. (1998). Applicants' reactions to the fairness of selection procedures: The effect of positive rule violations and time of measurement. *Journal of Applied Psychology, 83*, 3–16.

Ployhart, R. E., Ryan, A. M., & Bennett, M. (1999). Explanations for selection decisions: Applicants' reactions to informational and sensitivity features of explanations. *Journal of Applied Psychology, 84*, 87–106.

Premack, S. L., & Wanous, J. P. (1985). A meta-analysis of realistic job preview experiments. *Journal of Applied Psychology, 70*, 706–719.

Pulver, U. (1975). Die Krise der psychologischen Diagnostik – eine Koartationskrise. *Schweizerische Zeitschrift für Psychologie, 34*, 212–221.

Pulver, U., Lang, A. & Schmid, F. W. (Hrsg.). (1978). *Ist Psychodiagnostik verantwortbar? Wissenschaftler und Praktiker diskutieren Anspruch, Möglichkeiten und Grenzen psychologischer Erfassungsmittel.* Bern: Huber.

Rafaeli, A. (1999). Pre-employment screening and applicants' attitudes toward an employment opportunity. *The Journal of Social Psychology, 139*, 700–712.

Rauchfleisch, U. (1982). *Nach bestem Wissen und Gewissen. Die ethische Verantwortung in Psychologie und Psychotherapie.* Göttingen: Verlag für Medizinische Psychologie im Verlag Vandenhoeck & Ruprecht.

Ree, M. J., & Earles, J. A. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science, 1*, 86–89.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than g. *Journal of Applied Psychology, 79*, 518–524.

Remmers, H. H., Leidy, T. R., Starry, A. R., Shuman, D. L., & Tesser, A. (1966, April). High school students' attitudes on two controversial issues: War in Southeast Asia and the use of personality and ability tests. *Purdue Opinion, Panel, Report No. 77*.

Richman-Hirsch, W. L., Olson-Buchanan, J. B., & Drasgow, F. (2000). Examining the impact of administration medium on examinee perceptions and attitudes. *Journal of Applied Psychology, 85*, 880–887.

Robertson, I. T., Iles, P. A., Gratton, L., & Sharpley, D. (1991). The impact of personnel selection and assessment methods on candidates. *Human Relations, 44*, 963–982.

Robertson, I. T., & Kandola, R. S. (1982). Work sample tests: Validity, adverse impact and applicant reactions. *Journal of Occupational Psychology, 55*, 171–183.

Rolland, F., & Steiner, D. D. (2007). Test-taker reactions to the selection process: Effects of outcome favorability, explanations, and voice on fairness perceptions. *Journal of Applied Social Psychology, 37*, 2800–2836.

Rosse, J. G., Miller, J. L., & Stecher, M. D. (1994). A field study of job applicants' reactions to personality and cognitive ability testing. *Journal of Applied Psychology, 79*, 987–992.

Rosse, J. G., Ringer, R. C., & Miller, J. L. (1996). Personality and drug testing: An exploration of the perceived fairness of alternatives to urinalysis. *Journal of Business and Psychology, 10*, 459–475.

Runge, T. (1996). Studie zur Sozialen Validität. In Arbeitskreis Assessment Center e.V. (Hrsg.), *Assessment Center als Instrument der Personalentwicklung. Schlüsselkompetenzen, Qualitätsstandards, Prozessoptimierung.* Reihe Assessment Center, Band 3 (S. 286–297). Hamburg: Windmühle.

Ryan, A. M., & Chan, D. (1999). Perceptions of the EPPP: How do licensure candidates view the process? *Professional Psychology: Research and Practice, 30*, 519–530.

Ryan, A. M., Greguras, G. J., & Ployhart, R. E. (1996). Perceived job relatedness of physical ability testing for firefighters: Exploring variations in reaction. *Human Performance, 9*, 219–240.

Ryan, A. M., & Ployhart, R. E. (2000). Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *Journal of Management, 26*, 565–606.

Ryan, A. M., Ployhart, R. E., & Greguras, G. J. (1998). Test preparation programs in selection contexts: Self-selection and program effectiveness. *Personnel Psychology, 51*, 599–621.

Ryan, A. M., Ployhart, R. E., Greguras, G. J., & Schmit, M. J. (1997, April). *Predicting applicant withdrawal from applicant attitudes.* Paper presented at the 12th Annual Conference of the Society for Industrial and Organizational Psychology, St. Louis, MO.

Ryan, A. M., Sacco, J. M., McFarland, L. A., & Kriska, S. D. (2000). Applicant self-selection: Correlates of withdrawal from a multiple hurdle process. *Journal of Applied Psychology, 85*, 163–179.

Ryan, A. M., & Sackett, P. R. (1987). Pre-employment honesty testing: Fakability, reactions of test takers, and company image. *Journal of Business and Psychology, 1*, 248–256.

Rynes, S. L. (1991). Recruitment, job choice, and post-hire consequences: A call for new research directions. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed, Vol. 2, pp. 399–444). Palo Alto, CA: Consulting Psychologists Press.

Rynes, S. L. (1993). Who's selecting whom? Effects of selection practices on applicant attitudes and behavior. In N. Schmitt & W. Borman (Eds.), *Personnel selection in organizations* (pp. 240–274). San Francisco, CA: Jossey-Bass.

Rynes, S. L., Bretz, R. D., & Gerhart, B. (1991). The importance of recruitment in job choice: A different way of looking. *Personnel Psychology, 44*, 487–521.

Rynes, S. L., & Connerley, M. L. (1993). Applicant reactions to alternative selection procedures. *Journal of Business and Psychology, 7*, 261–277.

Rynes, S. L., & Miller, H. E. (1983). Recruiter and job influences on candidates for employment. *Journal of Applied Psychology, 68*, 147–154.

Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., & de Fruyt, F. (2003). International validity generalization of GMA and cognitive abilities: A European Community meta-analysis. *Personnel Psychology, 56*, 573–605.

Sanchez, R. J., Truxillo, D. M., & Bauer, T. N. (2000). Development and examination of an expectancy-based measure of test-taking motivation. *Journal of Applied Psychology, 85*, 739–750.

Saunders, D. M. (Ed). (1992). *New approaches to employee management: Fairness in employee selection* (Vol. 1). Greenwich, CT: JAI Press.

Schaubroeck, J., May, D. R., & Brown, F. W. (1994). Procedural justice explanations and employee reactions to economic hardship: A field experiment. *Journal of Applied Psychology, 79*, 455–460.

Schinkel, S., van Dierendonck, D., & Anderson, N. (2004). The impact of selection encounters on applicants: An experimental study into feedback effects after a negative selection decision. *International Journal of Selection and Assessment*, *12*, 197–205.

Schleicher, D. J., Venkataramani, V., Morgeson, F. P., & Campion, M. A. (2006). So you didn't get the job … Now what do you think? Examining opportunity-to-perform fairness perceptions. *Personnel Psychology, 59*, 559–590.

Schmid, K. (1971). *Psychologische Testverfahren im Personalbereich. Eine Darstellung ihrer rechtlichen Problematik für Personalleiter, Psychologen und Juristen.* Köln: Dr. E. W. Müssener-Verlag.

Schmidt, F. L., Greenthal, A. L., Hunter, J. E., Berner, J. G., & Seaton, F. W. (1977). Job sample vs. paper-and-pencil traders and technical tests: Adverse impact and examinee attitudes. *Personnel Psychology, 30*, 187–197.

Schmidt, F. L., & Hunter, J. E. (1998a). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schmidt, F. L. & Hunter, J. E. (1998b). Messbare Personenmerkmale: Stabilität, Variabilität und Validität zur Vorhersage zukünftiger Berufsleistung und berufsbezogenen Lernens. In M. Kleinmann & B. Strauss (Hrsg.), *Potentialfeststellung und Personalentwicklung* (S. 15–43). Göttingen: Verlag für Angewandte Psychologie.

Schmit, M. J., & Ryan, A. M. (1992). Test-taking dispositions: A missing link? *Journal of Applied Psychology, 77*, 629–637.

Schmit, M. J., & Ryan, A. M. (1997). Applicant withdrawal: The role of test-taking attitudes and racial differences. *Personnel Psychology, 50*, 855–876.

Schmitt, N., & Coyle, B. W. (1976). Applicant decisions in the employment interview. *Journal of Applied Psychology, 61*, 184–192.

Schmitt, N., & Gilliland, S. W. (1992). Beyond differential prediction: Fairness in selection. In D. M. Saunders (Ed.), *New approaches to employee management: Fairness in employee selection* (Vol. 1, pp. 21–46). Greenwich, CT: JAI Press.

Schmitt, N., Gilliland, S. W., Landis, R. S., & Devine, D. (1993). Computer-based testing applied to selection of secretarial applicants. *Personnel Psychology, 46*, 149–165.

Schreurs, B. (2007). *From post- to pretest applicant reactions. The effect of applicant selection expectations and perceptions on organizational attractiveness.* Unpublished doctoral dissertation, University of Leuven, Leuven, Belgium.

Schreurs, B., Derous, E., Proost, K., Notelaers, G., & De Witte, K. (2008). Applicant selection expectations: Validating a multidimensional measure in the military. *International Journal of Selection and Assessment, 16*, 170–176.

Schuler, H. (1990). Personenauswahl aus der Sicht der Bewerber: Zum Erleben eignungsdiagnostischer Situationen. *Zeitschrift für Arbeits- und Organisationspsychologie, 34*, 184–191.

Schuler, H. (1993a). Social validity of selection situations: A concept and some empirical results. In H. Schuler, J. L. Farr, & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 11–26). Hillsdale, NJ: Erlbaum.

Schuler, H. (1993b). Is there a dilemma between validity and acceptance in the employment interview? In B. Nevo & R. S. Jäger (Eds.), *Educational and psychological testing: The test taker's outlook* (pp. 239–250). Toronto: Hogrefe & Huber.

Schuler, H. (1998). *Psychologische Personalauswahl. Einführung in die Berufseignungsdiagnostik* (2. unveränd. Aufl.). Göttingen: Verlag für Angewandte Psychologie.

Schuler, H. (2002). *Das Einstellungsinterview.* Göttingen: Hogrefe.

Schuler, H. & Funke, U. (1987). Tests sollen keine Tortur sein. *Psychologie Heute, 14*(9), 58–65.

Schuler, H., Hell, B., Trapmann, S., Schaar, H. & Boramir, I. (2007). Die Nutzung psychologischer Verfahren der externen Personalauswahl in deutschen Unternehmen. Ein Vergleich über 20 Jahre. *Zeitschrift für Personalpsychologie, 6*, 60–70.

Schuler, H. & Stehle, W. (1983). Neuere Entwicklungen des Assessment-Center-Ansatzes – beurteilt unter dem Aspekt der sozialen Validität. *Psychologie und Praxis. Zeitschrift für Arbeits- und Organisationspsychologie, 27*, 33–44.

Schuler, H. & Stehle, W. (1985). Soziale Validität eignungsdiagnostischer Verfahren: Anforderungen für die Zukunft. In H. Schuler & W. Stehle (Hrsg.), *Organisationspsychologie und Unternehmenspraxis. Perspektiven der Kooperation* (S. 133–138). Stuttgart: Verlag für Angewandte Psychologie.

Schweizerische Gesellschaft für Psychologie (1975). Jahresversammlung 1975. Symposium Krise der Diagnostik. *Schweizerische Zeitschrift für Psychologie, 34*, 205–249.

Schweizerische Gesellschaft für Psychologie (1976). Krise der Diagnostik. Fortsetzung der Diskussion. *Schweizerische Zeitschrift für Psychologie, 35*, 49–61.

Shaw, J. C., Wild, E., & Colquitt, J. A. (2003). To justify or excuse? A meta-analytic review of the effects of explanations. *Journal of Applied Psychology, 88*, 444–458.

Sheppard, B. H., & Lewicki, R. J. (1987). Toward general principles of managerial fairness. *Social Justice Research, 1*, 161–176.

Sichler, R. (1989). Das Erleben und die Verarbeitung eines Assessment-Center-Verfahrens. Ein empirischer Beitrag zur "Sozialen Validität" eignungsdiagnostischer Situationen. *Zeitschrift für Arbeits- und Organisationspsychologie, 33*, 139–145.

Sieber, G. (1969). *Achtung Test*. Stuttgart: Deutsche Verlags-Anstalt.

Simmons, D. D. (1968). Invasion of privacy and judged benefit of personality-test inquiry. *The Journal of General Psychology, 79*, 177–181.

Singer, M. (1990). Determinants of perceived fairness in selection practices: An organizational justice perspective. *Genetic, Social, and General Psychology Monographs, 116*, 477–494.

Singer, M. S. (1992). Procedural justice in managerial selection: Identification of fairness determinants and associations of fairness perceptions. *Social Justice Research, 5*, 49–70.

Singer, M. S. (1993). *Fairness in personnel selection*. Aldershot, UK: Avebury.

Smither, J. W., Reilly, R. R., Millsap, R. E., Pearlman, K., & Stoffey, R. W. (1993). Applicant reactions to selection procedures. *Personnel Psychology, 46,* 49–76.

Society for Industrial and Organizational Psychology (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Spörli, S. (1978). *Kritische Theorie diagnostischer Praxis – dargestellt am Beispiel Verkehrspsychologie.* Bern: Huber.

Steiner, D. D., & Gilliland, S. W. (1996). Fairness reactions to personnel selection techniques in France and the United States. *Journal of Applied Psychology, 81,* 134–141.

Stephenson, W. (1953). *The study of behavior: Q-technique and its methodology.* Chicago, IL: The University of Chicago Press.

Stinglhamber, F., Vandenberghe, C. & Brancart, S. (1999). Les réactions des candidates envers les techniques de sélection du personnel: Une étude dans un contexte francophone. *Le Travail Humain, 62*, 347–361.

Stoffey, R. W., Millsap, R. E., Smither, J. W. & Reilly, R. R. (1991, April). *The influence of selection procedures on attitudes about the organization and job pursuit intentions.* Paper presented at the 6th Annual Conference of the Society for Industrial and Organizational Psychology, St. Louis, MO.

Stoll, F. (1977). Zur Abhängigkeit des Eignungsdiagnostikers und des Probanden: Lösungsvorschläge. In J. K. Triebe & E. Ulich (Hrsg.), *Beiträge zur Eignungsdiagnostik* (S. 203–213). Bern: Huber.

Stone, E. F., & Stone, D. L. (1990). Privacy in organizations: Theoretical issues, research findings, and protection mechanisms. In G. R. Ferris & K. M. Rowland (Eds.), *Research in personnel and human resource management* (Vol. 8, pp. 349–411). Greenwich, CT: JAI Press.

Stone-Romero, E. F., Stone, D. L., & Hyatt, D. (2003). Personnel selection procedures and invasion of privacy. *Journal of Social Issues, 59*, 343–368.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams, R. M. (1949). *The American soldier: Adjustment during army life* (Vol. 1). Princeton, NJ: Princeton University Press.

Strauss, A., & Corbin, J. (1997). *Grounded theory in practice*. Thousand Oaks, CA: Sage.

Streicher, B., Jonas, E., Maier, G. W., Frey, D., Woschée, R., & Wassmer, B. (2008). Test of the construct and criteria validity of a German measure of organizational justice. *European Journal of Psychological Assessment, 24*, 131–139.

Tachler, E. (1983). *Empirischer Beitrag zur Erforschung der sozialen Validität.* Unveröff. Bericht, Universität Hohenheim.

Taylor, M. S., & Bergmann, T. J. (1987). Organizational recruitment activities and applicants' reactions at different stages of the recruitment process. *Personnel Psychology, 40*, 261–285.

Teel, K. S., & Dubois, H. (1983). Participants' reactions to assessment centers. *Personnel Administrator, 28*, 85–91.

Tesser, A., & Leidy, T. R. (1968). Psychological testing through the glass of youth. *American Psychologist, 23*, 381–384.

Thibaut, J., & Walker, L. (1975). *Procedural justice: A psychological analysis.* Hillsdale, NJ: Erlbaum.

Thornton, G. C., III. (1993). The effect of selection practices on applicants' perceptions of organizational characteristics. In H. Schuler, J. L. Farr, & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 57–69). Hillsdale, NJ: Erlbaum.

Thornton, G. C., III, & Byham, W. C. (1982). *Assessment centers and managerial performance*. San Diego, CA: Academic Press.

Thorsteinson, T. J., Palmer, E. M., Wulff, C., & Anderson, A. (2004). Too good to be true? Using realism to enhance applicant attraction. *Journal of Business and Psychology, 19*, 125–137.

Thorsteinson, T. J., & Ryan, A. M. (1997). The effect of selection ratio on perceptions of the fairness of a selection test battery. *International Journal of Selection and Assessment, 5*, 159–168.

Triebe, J. K. & Ulich, E. (Hrsg.). (1977). *Beiträge zur Eignungsdiagnostik.* Bern: Huber.

Truxillo, D. M., & Bauer, T. N. (1999). Applicant reactions to test score banding in entry-level and promotional contexts. *Journal of Applied Psychology, 84*, 322–339.

Truxillo, D. M., Bauer, T. N., Campion, M. A., & Paronto, M. E. (2002). Selection fairness information and applicant reactions: A longitudinal field study. *Journal of Applied Psychology, 87*, 1020–1031.

Truxillo, D. M., Bauer, T. N., & Sanchez, R. J. (2001). Multiple dimensions of procedural justice: Longitudinal effects on selection system fairness and test-taking self-efficacy. *International Journal of Selection and Assessment, 9*, 336–349.

Truxillo, D. M., Bodner, T. E., Bertolino, M., Bauer, T. N., & Yonce, C. A. (2009). Effects of explanations on applicant reactions: A meta-analytic review. *International Journal of Selection and Assessment, 17*, 346–361.

Truxillo, D. M., Steiner, D. D., & Gilliland, S. W. (2004). The importance of organizational justice in personnel selection: Defining when selection fairness really matters. *International Journal of Selection and Assessment, 12*, 39–53.

Tyler, T. R., & Bies, R. J. (1990). Beyond formal procedures: The interpersonal context of procedurale justice. In J. S. Carroll (Ed.), *Applied social psychology and organizational settings* (pp. 77–98). Hillsdale, NJ: Erlbaum.

Uniform Guidelines on Employee Selection Procedures, 45 Fed. Reg. 74,676-74,677 (1966, 1978).

Van den Bos, K., Bruins, J., Wilke, H. A. M., & Dronkert, E. (1999). Sometimes unfair procedures have nice aspects: On the psychology of the Fair Process Effect. *Journal of Personality and Social Psychology, 77*, 324–366.

Van Vianen, A. E. M., Taris, R., Scholten, E., & Schinkel, S. (2004). Perceived fairness in personnel selection: Determinants and outcomes in different stages of the assessment procedure. *International Journal of Selection and Assessment, 12*, 149–159.

Wanous, J. P. (1992). *Organizational entry: Recruitment, selection, orientation, and socialization of newcomers* (2nd ed.). Reading, MA: Addison-Wesley.

Weitz, J. (1956). Job expectancy and survival. *Journal of Applied Psychology, 40*, 245–247.

Westhoff, K. (Hrsg.). (2006). *Nutzen der DIN 33430. Praxisbeispiele und Checklisten.* Lengerich: Pabst Science Publishers.

Westhoff, K., Hellfritsch, L. J., Hornke, L. F., Kubinger, K. D., Lang, F., Moosbrugger, H., Püschel, A. & Reimann, G. (Hrsg.). (2004). *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430.* Lengerich: Pabst Science Publishers.

Westmeyer, H. (2004). Die sogenannte Krise in der psychologischen Diagnostik. Erinnerungen an die 70er Jahre des 20. Jahrhunderts. *Diagnostica, 50*, 10–16.

6. Konstruktion des Leadership-Fragebogens

6.1 Anforderungen an militärische Kader und Bestimmung der Dimensionen des Leadership-Fragebogens

„Eine Arbeits- und Anforderungsanalyse ... sollte die Basis einer Eignungsbeurteilung sein" (DIN, 2002, S. 12), schreibt die DIN 33430 „Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen" vor. Dies gilt entsprechend auch für die Entwicklung eines Tests für den Einsatz im Rahmen der Personalselektion: So ist hier in einem ersten Schritt eine Arbeits- und Anforderungsanalyse durchzuführen, in welcher die wichtigsten Arbeitsinhalte und das für deren erfolgreiche Bewältigung erforderliche Wissen, die Fertigkeiten und Fähigkeiten erhoben werden (Wheaton & Whetzel, 2007). Nur wenn bekannt ist, welche Anforderungen ein bestimmter Arbeitsplatz an einen Mitarbeiter stellt, wird man in der Lage sein, einen geeigneten Mitarbeiter gezielt auswählen zu können (Kanning & Holling, 2002). Und nur bei einer guten Passung der erfolgsrelevanten Anforderungen einer Arbeitsstelle mit den anlässlich der Selektion erhobenen Merkmalen des Bewerbers – also bei einer hohen Inhaltsvalidität der eingesetzten Instrumente – schafft man gute Voraussetzungen für eine hohe Aussagekraft des Selektionsverfahrens (Rose & Baydoun, 1995). Zudem ist – wie ich in Kapitel 5 aufgezeigt habe – der erlebte Zusammenhang zwischen den Inhalten der Selektionsverfahren und denjenigen der Arbeitsstelle der wichtigste Aspekt für das Empfinden von Fairness im Selektionsprozess (z. B. Ryan & Ployhart, 2000; Schleicher, Venkataramani, Morgeson & Campion, 2006). Die beiden nachfolgend aufgeführten Definitionen beschreiben in knapper Form den Inhalt von Arbeits- und Anforderungsanalysen:

> *Job analysis* is the systematic process of discovery of the nature of a job by dividing it into smaller units, where the process results in one or more written products (Brannick & Levine, 2002, S. 9).
>
> Job analysis is a study of what a jobholder does on the job, what must be known in order to do it, what resources are used in doing it, and perhaps the conditions under which it is done (Guion, 1998, S. 58).

Eckardt und Schuler (1992; siehe auch Kanning, 2004) unterscheiden bei Anforderungsanalysen drei methodische Zugänge:

> *Erfahrungsgeleitet-intuitive Methode*: In Workshops erstellen Experten Listen von tätigkeitsrelevanten Eigenschaften und stützen sich dabei auf ihren Erfahrungsschatz, auf vorhandene Informationen und Dokumentati-

onen zur Tätigkeit und deren Anforderungen. Dieses Vorgehen ist zwar sehr ökonomisch, dafür ist die Qualität des Arbeitsergebnisses zum Teil fragwürdig und zudem stark von der Kompetenz der Experten abhängig (Hell, Ptok & Schuler, 2007). Die Methode ist jedoch sehr gut geeignet, um einen ersten Eindruck oder Anhaltspunkte für weitere Untersuchungen zu gewinnen (Reimann, 2005).

Arbeitsanalytisch-empirische Methode: Die Analyse erfolgt hier durch den Einsatz teil- oder vollstandardisierter Erhebungsinstrumente, wie Fragebogen, Checklisten oder Arbeitsanalyseverfahren, anhand welcher konkrete Arbeitsplätze untersucht werden. Auf der Grundlage der erhobenen Merkmale der Tätigkeit lassen sich Anforderungen formulieren, die einen direkten Bezug zu den ermittelten Personenmerkmalen ermöglichen.

Personenbezogen-empirische Methode: Die Anforderungen bestimmen sich bei dieser Methode aus den statistischen Zusammenhängen zwischen Personenmerkmalen und Kriterien wie der erbrachten Leistung oder der Arbeitszufriedenheit. Die Schwierigkeit dieses sehr aufwändigen Ansatzes liegt darin, eine erschöpfende Liste möglicher Prädiktoren der interessierenden beruflichen Leistung zu erstellen.

Auf die verschiedenen konkreten Instrumente und Vorgehensweisen der Arbeits- und Anforderungsanalyse gehe ich an dieser Stelle nicht ein. Übersichten dazu sind in jedem Lehrbuch zur Personaldiagnostik und –selektion und in den entsprechenden Monografien zu finden (z. B. Brannick & Levine, 2002; Gael, 1983, 1988; Harvey, 1991; McCormick, 1979). Es lässt sich nicht a priori feststellen, welche dieser Vorgehensweisen und welcher der oben aufgeführten Ansätze am besten geeignet ist. In den *Principles for the Validation and Use of Personnel Selection Procedures* der Society for Industrial and Organizational Psychology (2003) steht dazu:

> There is no single approach that is the preferred method for the analysis of work. The analyses used in a specific study of work are a function of the nature of work, current information about the work, the organizational setting, the workers themselves, and the purpose of the study. (S. 11)

Zudem sind die Bedürfnisse und Ziele der Organisation bei der Auswahl der Analysemethode zu berücksichtigen. Der Detaillierungsgrad der durchzuführenden Analysen hängt auch massgeblich von eventuell schon vorliegenden Informationen oder Studien zur Tätigkeit ab. Da es hier um die Entwicklung eines Persönlichkeitstests geht, muss ich die Anforderungen auch nur bezogen auf Persönlichkeitseigenschaften beschreiben und kann eventuell notwendiges Wissen oder spezielle Fertigkeiten, welche in den meisten Fällen truppengattungs- und

funktionsspezifisch und deshalb nicht generalisierbar sind, unberücksichtigt lassen.

Im nachfolgenden Text beschreibe ich die Vorgehensweise bei der Bestimmung der Anforderungen an unteres Kader der Schweizer Armee, anhand welcher ich die Dimensionen des Leadership-Fragebogens ableitete.

Im Konzept zu den psychologischen Aspekten der Rekrutierung A XXI führte ich beim Test zur Erfassung der sozialen Kompetenz I – den Test bezeichnen wir in der heutigen Form als Leadership-Fragebogen – die Dimensionen Teamfähigkeit, Konfliktfähigkeit, Frustrationstoleranz und Dominanzstreben auf (siehe auch Bühler Ruedin & Selk, 2001). Diese erschienen mir nach der Durchsicht der wichtigsten dazu vorliegenden Veröffentlichungen (Annen, 2000; Hoenle, 1996; Schweizer Armee, 1995; Stadelmann, 1998; Steiger, 1999) und auf Grund eigener Erfahrungen zentrale Persönlichkeitsmerkmale eines Unteroffiziers der Schweizer Armee zu sein. Die definitive Auswahl, Definition und Benennung der Dimensionen des Leadership-Fragebogens legte ich zu Beginn der Testkonstruktion fest. Dazu befasste ich mich nochmals eingehend mit den Arbeiten zur Führungskultur in der Schweizer Armee von Hoenle (1996) und zur förderwirksamen Beurteilung von Annen (2000) und zog bisher nicht berücksichtige Quellen hinzu (Fuhrer, 1985; Gonin, 1993; Schweizer Armee, 1997; Steiger & Annen, 1997; Zollinger, 1997). Nachfolgend stelle ich die für die Wahl der Dimensionen des Leadership-Fragebogens relevanten Aussagen aus den empirisch durchgeführten Arbeiten von Hoenle und Annen dar und beschreibe, wie der militärische Vorgesetzte im Dienstreglement der Schweizer Armee charakterisiert wird.

Hoenle (1996) interviewte für seine Bestandesaufnahme der militärischen Führungskultur 21 Berufsoffiziere der Schweizer Armee zu ihren persönlichen Erfahrungen mit Führung. Auf der Grundlage der qualitativen Auswertung der Interviews leitet er Grundannahmen zur Führeridentität, zum Führungsprozess und zu den Führungsbedingungen ab und beschreibt vier Führertypen. Die Führeridentität charakterisiert er anhand von sechs Dimensionen der Führungskultur in der Schweizer Armee: (In Klammern führe ich die entsprechenden Persönlichkeitseigenschaften auf.)

Frühe Prägung durch Leitfiguren

Die in jungen Jahren gemachten persönlichen Erfahrungen mit Vorgesetzten oder der eigenen Führungstätigkeit – zum Beispiel als Jugendgruppenleiter – beeinflussen die nachfolgende Kaderlaufbahn massgeblich. Die ersten militärischen Vorgesetzten können dabei als Leitfiguren wirken, denen man nacheifern möchte. Motivierend wirkt auch, wenn einem schon als Rekrut oder Soldat zeitweise eine Führungsfunktion übertragen wird. (Begeisterungsfähigkeit)

Persönlichkeit und Kompetenz

Damit ihn die Unterstellten akzeptieren, muss ein militärischer Vorgesetzter persönlich, fachlich und führungsmässig überzeugen und nicht auf Grund seiner institutionell verliehenen Gradautorität. Der Chef muss die Hintergründe eines Befehls transparent machen können, um die Mannschaft zu überzeugen. Dies setzt Fachwissen und eine gewisse Intelligenz voraus. Die wichtigsten Persönlichkeitseigenschaften sind Eigenständigkeit und Selbstsicherheit, um Befehle durchsetzen zu können, aber auch um die eigenen Schwächen zu erkennen und dazu zu stehen. (Kommunikationsfähigkeit, Fachwissen, Intelligenz, Eigenständigkeit, Selbstsicherheit, Durchsetzungsfähigkeit, Selbstreflexion)

Profil und Konturen

Als militärisches Kadermitglied benötigt man klare Konturen und muss sich in Abgrenzung zu anderen profilieren. Dies erreicht man durch unkonventionelles Vorgehen, durch Abgrenzung gegenüber seinen Vorgesetzten, durch das unmissverständliche Darlegen des eigenen Standpunktes oder durch ein sich Exponieren, indem man zum Beispiel die Zielerreichung durchsetzt. (Unabhängigkeit, Selbständigkeit, Selbstbewusstsein, Durchsetzungsfähigkeit)

Pflichtgefühl und Ehrgeiz

Führen bedeutet vor allem die Übernahme von Verantwortung für sich und die Unterstellten. Dazu muss man ein entsprechendes Pflichtgefühl entwickeln, sich zum Wohle des Ganzen einsetzen, loyal sein und dabei seine persönlichen Motive und Wünsche zurückstellen. Führungspersönlichkeiten zeichnen sich auch durch Ehrgeiz, Disziplin und eine sehr hohe Leistungsmotivation aus: Man will die übertragene Arbeit möglichst gut erledigen. (Verantwortungsübernahme, Pflichtbewusstsein, Engagement, Loyalität, Ehrgeiz, Disziplin, Leistungsmotivation)

Einsamkeit

Der Vorgesetzte fällt seine Entscheide selbständig und trägt auch die damit verbundene Verantwortung alleine. Er biedert sich bei den Unterstellten nicht an. (Selbständigkeit, Unabhängigkeit)

Bewirker und Beweger

Der Vorgesetzte lenkt aktiv die Geschehnisse, indem er präsent ist, kontrolliert und direkt Einfluss nimmt und so bei den Unterstellten einen Gehorsam ohne Zögern bewirkt. Seine Handlungen sind dabei konsequent auf den Erfolg ausgerichtet und er setzt einfache, klare und bedeutungsvolle Ziele. Der Vorgesetzte erhält so die Überzeugung, Kontrolle ausüben zu können. (Beeinflussungsverhalten, Zielorientierung, Kontrollüberzeugung)

Das von Annen (2000) entwickelte förderwirksame Beurteilungsverfahren für Milizkader der Schweizer Armee stellte lange Zeit die wichtigste und umfassendste Arbeit zu den Anforderungen an unteres Milizkader dar. Die Grundlage dazu bildeten halbstrukturierte Interviews mit Schulkommandanten zum Thema Qualifikation und eine gross angelegte schriftliche Befragung von Instruktoren und Offiziersschülern nach der Wichtigkeit verschiedener Qualifikationskriterien. Das auf diesen Ergebnissen aufbauende Qualifikationssystem enthält eine Auflistung verhaltensbezogener Eigenschaften, welche Annen in die Kategorien Selbst- und Sozialkompetenz und Führungsverhalten unterteilt. Damit entstand ein Anforderungsprofil für Milizoffiziere, welches ich im Folgenden in vereinfachter Form darstelle:

Selbst- und Sozialkompetenz

> *Persönliche Grundhaltung*: verlässlich, geradlinig, loyal, selbständig, eigenverantwortlich, kritikfähig, lösungsorientiert, initiativ, engagiert, belastbar
>
> *Geistige Fähigkeiten*: fasst rasch und vollständig auf, erfasst und beurteilt wesentliche Zusammenhänge, transferiert Wissen, lernfähig
>
> *Soziales Verhalten*: geradliniger und kollegialer Umgang, offen, verständnisvoll, fürsorglich, konfliktfähig, teamfähig

Führungsverhalten

> *Fähigkeit als Führer*: identifiziert sich mit der Aufgabe, zielstrebig, beharrlich, überzeugend, setzt sich durch, belastbar, zuverlässig, konzentriert, überzeugt, wirkt mitreissend, qualifiziert fair und seriös
>
> *Kommunikationsverhalten*: kommuniziert offen und direkt, hört aktiv zu, informiert regelmässig und stufengerecht, tritt natürlich auf
>
> *Führungstechnik*: befiehlt klar und situationsgerecht, gliedert Aufträge in Teilschritte, hält den Führungsrhythmus (Reihenfolge der Führungstätigkeiten) ein, kontrolliert, delegiert
>
> *Fähigkeit als Ausbilder*: setzt angemessene Ziele und verfolgt diese konsequent, vermittelt Ausbildungsinhalte empfängerorientiert, praxisbezogen und methodisch richtig

Im damals gültigen Dienstreglement 95 (Schweizer Armee, 1995) sind in den Ziffern 9 bis 17 die Führungsgrundsätze der Schweizer Armee aufgeführt. Im Anschluss an die Beschreibung der einzelnen Ziffern führe ich die entsprechenden Persönlichkeitseigenschaften auf:

Führung: Ausrichten des Handelns der Unterstellten auf das Erreichen eines Ziels. (Beeinflussungsverhalten, Durchsetzungsfähigkeit)

Führen durch Zielvorgabe: Die Kader geben das zu erreichende Ziel vor und lassen den Unterstellten bei der Umsetzung Handlungsfreiheit. Dieses Vorgehen verlangt von den Kadern Mut, Vertrauen und Respekt. (Wertschätzung der Unterstellten)

Mitdenken und Engagement: Führen durch Zielvorgabe verlangt von den Unterstellten aktives Mitdenken und selbständiges Handeln.

Verantwortung: Der Vorgesetzte trägt die Verantwortung für lagegerechte und zeitgerechte Aufträge. Er ist sich der Folgen bewusst und kontrolliert die Zielerreichung. Er berücksichtigt die Fähigkeiten der Unterstellten und ist für deren Wohl und Schutz verantwortlich. (Verantwortungsübernahme, Fürsorge)

Disziplin: Alle Armeeangehörigen stellen ihre eigenen Interessen und Wünsche zugunsten des Ganzen zurück und weisen eine hohe Leistungsmotivation auf. (Disziplin, Leistungsmotivation)

Information: Der Vorgesetzte informiert jederzeit und umfassend über seine Absicht. Zudem bemüht er sich, die für die Erfüllung seines Auftrages wichtigen Informationen zu erhalten. (Informations- und Kommunikationsfähigkeit)

Kommunikation: Durch laufende Kommunikation wird erreicht, dass sich alle mit dem Auftrag identifizieren. Zudem fördert sie das Vertrauen der Unterstellten in den Vorgesetzten. (Kommunikationsfähigkeit, Vertrauenswürdigkeit)

Vorbild: Autorität erhält der Vorgesetzte durch seine fachliche und persönliche Glaubwürdigkeit. Er wirkt als persönliches Vorbild, indem er Disziplin und Engagement vorlebt. (fachliche Kompetenz, Gewissenhaftigkeit, Disziplintiertheit, Engagement)

Zusammenhalt und Leistung: Die Kader achten die Unterstellten, vertrauen diesen und setzen sich für den Zusammenhalt und die Stärkung der Leistungskraft des Verbandes ein. (Wertschätzung der Unterstellten, Engagement)

Die von Hoenle und Annen aufgelisteten Anforderungen an militärisches Kader und die im Dienstreglement vorgeschriebenen Verhaltensweisen verwendeten wir als Ausgangsmaterial für die Bildung eines provisorischen Anforderungsprofils für militärische Führungskräfte, welches in Tabelle 6.1 dargestellt ist.

Tabelle 6.1
Provisorisches Anforderungsprofil für untere Kader der Schweizer Armee

Kognitive Leistungsfähigkeit	Intelligenz, Auffassungsgabe, Konzentrationsfähigkeit, erkennt Zusammenhänge, Lernfähigkeit, Lerntransfer, Fachwissen
Leistungsmotivation	Leistungsmotivation, Engagement, Initiative, Ehrgeiz, Lösungsorientierung, Zielorientierung, Identifikation, Begeisterungsfähigkeit
Belastbarkeit	Belastbarkeit
Gewissenhaftigkeit	Gewissenhaftigkeit, Zuverlässigkeit, Disziplin iertheit, Loyalität
Selbstsicherheit	Selbstsicherheit, Selbstbewusstsein, Kontrollüberzeugung, Unabhängigkeit, Selbständigkeit
Durchsetzungsfähigkeit	Durchsetzungsfähigkeit, Beeinflussungsverhalten, Beharrlichkeit, Geradlinigkeit
Kommunikationsfähigkeit	Kommunikationsverhalten, Informationsverhalten, Zuhörfähigkeit
Kontaktfähigkeit	Teamfähigkeit, Aufgeschlossenheit, Konfliktfähigkeit, Kritikfähigkeit, Vertrauenswürdigkeit, Wertschätzung, Fairness, Verständnisbereitschaft
Verantwortungsbewusstsein	Verantwortungsübernahme, Pflichtbewusstsein, Fürsorglichkeit, Eigenverantwortlichkeit, Selbstreflexion

Die kognitive Leistungsfähigkeit wird anlässlich der Rekrutierung mit einem Intelligenz- und einem Merkfähigkeitstest erfasst. Die Dimensionen Leistungsmotivation, Belastbarkeit und Gewissenhaftigkeit operationalisierten wir in einem traditionellen, likert-skalierten Persönlichkeits-Fragebogen, welcher zusätzlich noch die Dimensionen Extraversion, Teamfähigkeit und Entgegenkommen/Friedfertigkeit umfasst. Die Dimensionen Selbstsicherheit (Auftreten) und Kommunikationsfähigkeit werden anlässlich der Grundrekrutierung nicht mittels psychologischer Testverfahren oder Übungen überprüft. Die Rekrutierungsoffiziere beachten diese Aspekte jedoch anlässlich des persönlichen Zuteilungsgespräches mit dem Stellungspflichtigen. Die verbleibenden drei Dimensionen Durchsetzungsfähigkeit, Kontaktfähigkeit und Verantwortungsbewusstsein operationalisierten wir im Leadership-Fragebogen, was eine leichte Abweichung von den im Konzept aufgeführten Dimensionen bedeutet. In Tabelle 6.2 sind die Änderungen zusammen mit den Definitionen der drei definitiven Skalen des Leadership-Fragebogens aufgeführt. Für die ursprüngliche Dimension Teamfähigkeit wählten wir das etwas weiter gefasste Konzept Kontaktfähigkeit, in welchem der Aspekt der Konfliktfähigkeit enthalten ist. Die Dimension Frustrationstoleranz strichen wir auf Grund der Ähnlichkeit mit der Dimension Durchsetzungsfähigkeit – Weiterverfolgen der gefassten Absicht trotz Hindernissen – ersatzlos. Für die Dimension Dominanzstreben wählten wir den weniger negativ belasteten Begriff Durchsetzungsfähigkeit.

Tabelle 6.2
Auflistung der Dimensionen des Tests zur Erfassung der sozialen Kompetenz I und des Leadership-Fragebogens

Test zur Erfassung der soz. Kompetenz	Leadership-Fragebogen	Kurz-Definition
Teamfähigkeit	Kontaktfähigkeit	Kontaktfähigkeit ist als Sammelbegriff für das Interesse am Mitmenschen zu verstehen, was sich in Geselligkeit, Offenheit und Umgänglichkeit äussert.
Konfliktfähigkeit	–	
Frustrationstoleranz	–	
Dominanzstreben	Durchsetzungsfähigkeit	Durchsetzungsfähigkeit beschreibt die Eigenschaft, die eigenen Interessen gegenüber anderen zu wahren, um so Widerstände zu überwinden, welche durch andere Personen verursacht sind.
–	Verantwortungsbewusstsein	Verantwortungsbewusstsein zeichnet sich durch die Antizipation der Konsequenzen des eigenen Handelns und durch eine fürsorgliche Haltung gegenüber der Unterstellten aus.

Als 2009 der Chef der Armee dem Chef Personelles der Armee den Auftrag erteilte, die Kaderselektion in der Schweizer Armee zu vereinheitlichen, erstellte die aus Vertretern aller Lehrverbänden gebildete Arbeitsgruppe – insgesamt 16 Personen – als Grundlage für die nachfolgende Planung der Selektionsprozesse in drei Workshopgruppen ein Anforderungsprofil für die untere Führungsstufe. Dieses nach der erfahrungsgeleitet-intuitiven Methode der Anforderungsanalyse (Eckardt & Schuler, 1992) erstellte Anforderungsprofil ist in Tabelle 6.3 dargestellt. Bemerkenswert daran ist, dass sich drei der vier Anforderungsdimensionen mit den Skalen im Leadership-Fragebogen decken. Als zusätzliche Kategorie haben die Experten noch die Belastbarkeit / Leistungsbereitschaft gebildet.

Zeitgleich hatten wir vom Kommando Rekrutierung den Auftrag erhalten, Anforderungsprofile für Gruppenführer und Zugführer[1] zu erstellen. Dazu führten wir in fünf Rekruten- und Kaderschulen der Schweizer Armee Workshops mit insgesamt 22 Berufsoffizieren und –unteroffizieren durch, welche sich mit der Selektion und Ausbildung des unteren Kaders befassen. Wir setzten ein zweistufiges Verfahren ein, welches aus dem Sammeln und Diskutieren der Eigenschaften, Fähigkeiten und Kenntnisse einer Führungsperson der jeweiligen Kaderstufe und dem schriftlichen Notieren und anschliessenden Diskutieren erfolgskritischer Situationen und den dazugehörenden Verhaltensweisen mit der Critical Incident Technique bestand. Die Workshops dauerten ungefähr zwei Stunden und es nah-

[1] Zusätzlich erstellten wir auch noch Anforderungsprofile für Feldweibel und Fouriere.

men jeweils drei bis vier Experten daran teil. Die beiden wissenschaftlichen Mitarbeiterinnen protokollierten deren Aussagen und zeichneten das gesamte Gespräch zusätzlich auf.

Tabelle 6.3
Das von der Arbeitsgruppe Kaderselektion erstellte Anforderungsprofil für unteres Kader der Schweizer Armee

Teamfähigkeit, Kontakt- und Kommunikationsbereitschaft	arbeitet gerne mit Menschen zusammen; ist aufgeschlossen und humorvoll; integriert alle ins Team; ist hilfsbereit; bringt sich aktiv ein; hört zu, lässt sprechen; kommuniziert offen, direkt und ehrlich; bleibt bei Meinungsverschiedenheiten ruhig und korrekt; akzeptiert andere Lösungsvorschläge und Meinungen; stellt fest, dass andere Probleme haben; bietet Unterstützung an; nimmt Rücksicht; behandelt alle gleich
Verantwortungsbereitschaft	ist bereit Verantwortung für andere zu übernehmen; macht auf kritische Punkte aufmerksam; ist zeitgerecht; ist auftragstreu; ist zuverlässig; nutzt vorhandenen Spielraum vollständig aus; erkennt die Situation und den daraus entstehenden Handlungsbedarf; macht eine angepasste Lagebeurteilung
Durchsetzungsvermögen	setzt sich durch; kann sich Gehör verschaffen; entscheidet auch gegen Widerstände; bringt eigenen Standpunkt ein; bleibt beharrlich wo nötig; handelt konsequent und zielorientiert; stellt sich Herausforderungen; löst Probleme; ist ein Leadertyp; ist selbstbewusst und hat ein sicheres Auftreten
Belastbarkeit und Leistungsbereitschaft	hält der körperlichen Belastung stand; ist körperlich fit; ist psychisch belastbar; bleibt auch unter Druck sachlich, ausgeglichen und zuverlässig; handelt ruhig und überlegt; ist bei Überraschungen nicht überfordert; zeigt konstante Arbeitsleistung; geht die Arbeit aus eigenem Antrieb direkt und pragmatisch an; ist bereit für Mehrarbeit; ist sich nicht zu schade für ungeliebte Aufgaben; gibt bei Rückschlägen nicht auf; hat eine positive Grundeinstellung; reagiert auf Kritik gelassen und lösungsorientiert

Wie oben schon erwähnt, setzten sich die Workshops aus mündlichen und schriftlichen Elementen zusammen. Die Einstiegsaufgabe orientiert sich dabei an der für die Entwicklung von Anforderungsprofilen eingesetzten Methode der Setzung durch Experten (Kannheiser, 1995; Kanning, 2002), auch als *Subject Matter Expert Panels* (Cascio & Aguinis, 2005) oder *Subject Matter Expert Workshops* (Gatewood, Field & Barrick, 2008) bezeichnet. Wir forderten die an den Workshops teilnehmenden Berufsoffiziere und –unteroffiziere auf, in einer von uns geleiteten Diskussion Eigenschaften, Fähigkeiten und Fertigkeiten eines als gut oder sehr gut qualifizierten Funktionsinhabers zu nennen.

Im zweiten Teil des Workshops baten wir die Berufsmilitärs, wichtige, erfolgsrelevante oder häufig auftretende Situationen, welche sie selbst bei Kaderangehörigen beobachtet haben, auf vorbereitete Zettel zu notieren. Reihum stellten sie sodann ihre Situationen vor und erklärten, wie sich ein erfolgreicher

und ein nicht erfolgreicher Funktionsinhaber hier verhalten haben respektive würden. Die anderen Teilnehmer hatten die Möglichkeit, diese Aussagen zu kommentieren und zu ergänzen. Dieses Vorgehen orientiert sich an der *Critical Incident Technique* (Flanagan, 1954), welche als etabliertes Verfahren zur Erstellung von Anforderungsprofilen gilt (Jetter, 2008; Schuler, 2002) und den teilstandardisierten Methoden der Anforderungsanalyse zugeordnet wird (Reimann, 2005). Dabei beschränkt sich die Standardisierung auf die Vorgabe einiger weniger Fragen zur Erfassung der Bedingungen erfolgskritischen Verhaltens. Ziel ist es, anhand der Schilderungen von Experten über effektives und wenig effektives Verhalten bei der Ausübung der zu analysierenden Arbeitsstelle Definitionen von wichtigen beziehungsweise häufig auftretenden oder so genannt kritischen Ereignissen der jeweiligen Arbeitstätigkeit zu sammeln. Diese Datensammlung geschieht mittels Einzel- oder Gruppeninterviews (Workshop), per Fragebogen oder durch direkte Aufzeichnung (Schuler, 2002).

Das Vorgehen bei der Critical Incident Technique lässt sich in drei Phasen unterteilen (Anderson & Wilson, 1997; Höft & Schuler, 2005; Kanning, 2002; Koch, Kici, Strobel & Westhoff, 2006):

1. *Sammeln der kritischen Ereignisse*: Ungefähr 20 aktuelle beziehungsweise ehemalige Arbeitsplatzinhaber (Experten) schildern etwa zehn wichtige und/oder häufig auftretende Ereignisse aus dem Berufsalltag eines Stelleninhabers, die sie selbst erlebt oder beobachtet haben.

2. *Sammeln konkreter Verhaltensweisen*: Die Experten beschreiben danach für jede einzelne Situation, wie sich ein Stelleninhaber verhält, wenn er die Situation sehr gut meistert oder wenn er versagt. Anhand dieser realistischen Schilderung ergeben sich für jedes Ereignis mehrere konkrete positive und negative Verhaltensweisen.

3. *Extraktion der zugrunde liegenden Anforderungsdimensionen*: Durch die Kategorisierung der gesammelten Verhaltensweisen entsteht das Anforderungsprofil. Dazu stehen grundsätzlich zwei Methoden zur Verfügung: Bei der qualitativen Methode erstellen Personalfachleute oder Psychologen auf der Basis der Verhaltensweisen ein Kategoriensystem. Bei der quantitativen Methode stuft eine grosse Anzahl Personen mittels der in einem Fragebogen zusammengestellten Liste der Verhaltensweisen eine real existierende Person ein. Mittels einer Faktorenanalyse lassen sich sodann die Anforderungsdimensionen bestimmen.

Um aus dem Ausgangsmaterial zu den einzelnen Verhaltensweisen zu gelangen, führten wir den ersten Schritt der qualitativen Inhaltsanalyse nach Mayring (2008), eine zusammenfassende Inhaltsanalyse, durch. Hierzu haben

wir den aufgezeichneten Text paraphrasiert, generalisiert und schliesslich reduziert, indem wir ähnliche Verhaltensweisen zusammenfassten.

Für die dritte Phase der Critical Incident Technique, die Extraktion der Anforderungsdimensionen, wendeten wir das Verfahren des parallelen Sortierens (Marx & Läge, 1995) an. Anhand dieses Verfahrens lassen sich Ähnlichkeiten zwischen Objekten herausarbeiten, indem durch den Sortiervorgang der einzelnen Verhaltensweisen nach und nach ein Kategoriensystem gebildet wird. Sechs Psychologinnen und Psychologen, welche sich beruflich mit Personalselektion befassen, dienten uns als Experten. Sie erhielten pro Kaderfunktion die auf Kärtchen gedruckten Verhaltensweisen und eine Beschreibung des Vorgehens für die Bildung des Kategoriensystems (siehe Anhang 6.1). Aus den so entstandenen je sechs Anforderungsprofilen bildeten drei Projektmitarbeiterinnen anlässlich eines Workshops die beiden definitiven Anforderungsprofile und ordneten die Verhaltensweisen den einzelnen Dimensionen zu. Bei Uneinigkeit, welcher Dimension eine bestimmte Verhaltensweise zugeordnet werden soll, wurde so lange diskutiert, bis ein Konsens gefunden wurde. In Tabelle 6.4 sind die Dimensionen der beiden Anforderungsprofile im Überblick dargestellt. Ein Ausschnitt aus den jeweiligen Verhaltensweisen kann im Anhang 6.2 und 6.3 eingesehen werden.

Tabelle 6.4
Die Dimensionen der Anforderungsprofile für Gruppen- und Zugführer

Gruppenführer		Zugführer
	Analysefähigkeit	
	Organisations- und Planungsfähigkeit	
–		Allgemeinbildung
	physische & psychische Belastbarkeit	
	Leistungsbereitschaft & Engagement	
Gewissenhaftigkeit		Gewissenhaftigkeit & Loyalität
	Offenheit & Flexibilität	
	Selbstreflexion	
Teamfähigkeit		Kooperationsfähigkeit
Fürsorglichkeit		Einfühlungsvermögen
–		Konflikt- & Kritikfähigkeit
	Kommunikationsfähigkeit	
	Durchsetzungsfähigkeit	
	Verantwortungsübernahme	
Selbstsicherheit		Auftreten als Chef

Tabelle 6.5 zeigt eine Übersicht über den Verlauf der Kategorisierung der extrahierten Verhaltensweisen. Das darin aufgeführte Basis-Anforderungsprofil für

unteres Kader ist eine aus den Listen mit Verhaltensweisen der beiden Anforderungsprofile durch Expertenrating gebildete Synthese. Es soll anlässlich der Grundrekrutierung in den Rekrutierungszentren und in den ersten Wochen der Rekrutenschule zum Einsatz gelangen, wenn es festzustellen gilt, welche der Stellungspflichtigen respektive der Rekruten die Grundvoraussetzungen für die Übernahme einer Kaderfunktion in der Armee erfüllen.

Tabelle 6.5
Übersicht über die Anzahl der extrahierten Verhaltensweisen

	Gruppenführer	Zugführer	Basis-Anforderungsprofil
Anzahl extrahierter Verhaltensweisen (Total)	432	904	
Anzahl Verhaltensweisen (ohne Mehrfachnennungen)	219	336	112
Anzahl Verhaltensweisen im Anforderungsprofil	160	174	38
Anzahl Dimensionen des Anforderungsprofils	13	15	6

Zur Erstellung des Basis-Anforderungsprofils wählten wir von jeder Anforderungsdimension (ohne Allgemeinbildung) acht Verhaltensweisen und stellten sie in abwechselnder Reihenfolge in einem Fragebogen zusammen. Abbildung 6.1 zeigt einen Ausschnitt aus dem in Excel umgesetzten Fragebogen.

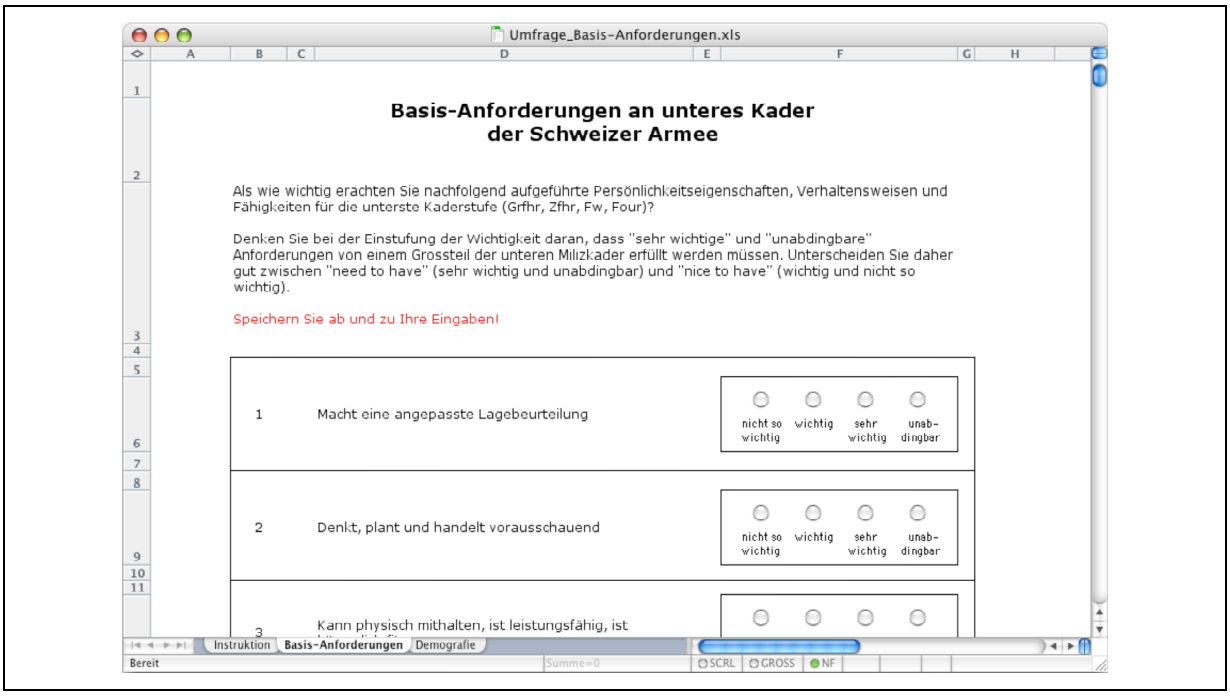


Abbildung 6.1 Ausschnitt aus dem Excel-File zur Befragung der Berufsmilitärs als Grundlage für die Erstellung des Basis-Anforderungsprofils.

Diesen Fragebogen stellte der Stellvertreter des Chefs Personelles der Armee Berufsoffizieren und –unteroffizieren zu, welche mit der Kaderselektion oder Kaderausbildung beauftragt sind. Diese stuften bei jeder der insgesamt 112 Verhaltensweisen deren Wichtigkeit für unteres Kader ein. Dazu stand ihnen eine vierstufige Antwortskala mit den Ausprägungen „nicht so wichtig", „wichtig", „sehr wichtig" und „unabdingbar" zur Verfügung. Sie erhielten folgende Anweisung:

> Als wie wichtig erachten Sie nachfolgend aufgeführte Persönlichkeitseigenschaften, Verhaltensweisen und Fähigkeiten für die unterste Kaderstufe (Grfhr, Zfhr, Fw, Four)?
>
> Denken Sie bei der Einstufung der Wichtigkeit daran, dass "sehr wichtige" und "unabdingbare" Anforderungen von einem Grossteil der unteren Milizkader erfüllt werden müssen. Unterscheiden Sie daher gut zwischen "need to have" (sehr wichtig und unabdingbar) und "nice to have" (wichtig und nicht so wichtig).

Insgesamt nahmen 63 Berufsmilitärs – davon eine Frau – an der Umfrage teil. Drei Teilnehmer schloss ich von den weiteren Berechnungen aus, da sie in einer oder mehreren Dimensionen zwei oder mehr fehlende Werte aufwiesen. Das Durchschnittsalter der 60 verbleibenden Teilnehmer beträgt 39.48 Jahre (*SD* = 7.75 Jahre, *Range* = 27 – 56 Jahre). Weitere Merkmale der Stichprobe habe ich in Tabelle 6.6 aufgeführt.

Tabelle 6.6
Arbeitsgebiet respektive Verband und militärischer Grad der Teilnehmer an der Datenerhebung für die Erstellung des Basis-Anforderungsprofils

Arbeitsgebiet / Verband	*n*	Militärischer Grad	*n*	%
Führungsstab der Armee	1	Adjutant Unteroffizier	7	11.7
Kompetenzzentren der Armee	6	Stabsadjutant	7	11.7
Lehrverband Infanterie	8	Hauptadjutant	5	8.3
Lehrverband Panzer / Artillerie	3	Chefadjutant	1	1.7
Lehrverband Genie / Rettung	5	Hauptmann	11	18.3
Lehrverband Logistik	24	Major / im Generalstab	9	15.0
Lehrverband Flieger	6	Oberstleutnant / im Generalstab	8	13.3
Lehrverband Fliegerabwehr	2	Oberst / im Generalstab	12	20.0
Lehrverband Führungsunterstützung	5			

Die anhand der Intraklassenkorrelation bestimmte Reliabilität der Einstufungen – die Beurteilerübereinstimmung – ist mit ICC_{unjust} = .18 (F(111, 4'995) = 15.05, p < .001) sehr gering, über alle Beurteilungen gemittelt mit $ICC_{unjust,MW}$ = .91

jedoch sehr gut (Wirtz & Caspar, 2002). Für die Bestimmung der Wichtigkeit der einzelnen Verhaltensweisen und Anforderungsdimensionen berechnete ich die entsprechenden Mittelwerte. (Die Einstufungen zu allen 112 Verhaltensweisen sind in den Anhängen 6.6 und 6.7 aufgeführt.) Wie der unten dargestellten Tabelle 6.7 zu entnehmen ist, stuften die Experten die Verhaltensweisen zur Verantwortungsübernahme durchschnittlich am wichtigsten ein (M = 3.08, SD = 3.48), diejenigen zur Teamfähigkeit am wenigsten wichtig (M = 2.47, SD = 3.45). Für die Bildung des Basis-Anforderungsprofils verwendete ich alle Verhaltensweisen – insgesamt 39 –, welche eine durchschnittliche Wichtigkeitseinschätzung von 2.85 oder mehr erhalten haben. Geplant war, nur Verhaltensweisen zur Bildung des Basis-Anforderungsprofils zu verwenden, welche die Berufsmilitärs durchschnittlich als mindestens sehr wichtig – dies entspricht dem Wert 3 – eingestuft haben. Auf diese Weise hätte sich das Anforderungsprofil jedoch nur aus 25 Verhaltensweisen zusammengesetzt, weshalb ich mich dazu entschlossen habe, den Grenzwert ein wenig tiefer zu legen. Diese 39 Verhaltensweisen gruppierte ich anhand der ursprünglichen Anforderungsdimensionen. Damit deren Anzahl überschaubar bleibt und pro definitive Dimension mehr als drei Verhaltensweisen enthalten sind, legte ich einzelne zusammen und ordnete einige Verhaltensweisen anderen Dimensionen zu. Das überarbeitete Basis-Anforderungsprofil ist in Tabelle 6.8 dargestellt.

Tabelle 6.7
Von den Berufsmilitärs eingestufte Wichtigkeit der Anforderungsdimensionen

Dimension	*Range*	*M*	*SD*	*M (1/8)*	*M Rang*	*SD*
Verantwortungsübernahme	18 – 30	24.60	3.48	3.08	21.50	16.96
Gewissenhaftigkeit & Loyalität	17 – 30	23.52	3.44	2.94	36.75	40.05
Auftreten als Chef / Selbstsicherheit	15 – 29	23.08	3.07	2.89	42.37	24.13
Planungs- & Organisationsfähigkeit	14 – 32	22.63	3.88	2.83	43.63	25.11
Fürsorglichkeit / Einfühlungsvermögen	14 – 30	22.45	3.29	2.81	48.63	35.92
Durchsetzungsfähigkeit	15 – 31	22.17	3.58	2.77	49.38	32.80
Leistungsbereitschaft & Engagement	15 – 30	22.08	3.47	2.76	48.00	27.61
Analysefähigkeit	13 – 27	21.23	3.23	2.65	63.75	30.04
Kommunikationsfähigkeit	12 – 28	21.15	3.35	2.64	64.63	28.83
Selbstreflexion	11 – 27	20.90	3.22	2.61	66.75	35.86
physische & psychische Belastbarkeit	14 – 27	20.52	3.27	2.56	73.88	31.79
Offenheit & Flexibilität	13 – 31	20.30	3.15	2.54	75.38	23.04
Konflikt- & Kritikfähigkeit	12 – 30	20.28	3.62	2.53	71.38	33.48
Teamfähigkeit	11 – 29	19.75	3.45	2.47	85.00	12.07

Anmerkung. *N = 60. Scoring der Wichtigkeitseinstufung:*
1 = nicht so wichtig; 2 = wichtig; 3 = sehr wichtig; 4 = unabdingbar.

Tabelle 6.8

Basis-Anforderungsprofil für unteres Kader der Schweizer Armee

Dimension	Verhaltensweisen	*Wichtigkeit*	
		M	*SD*
Gewissenhaftigkeit und Loyalität	ist ein Vorbild	3.64	.52
	ist zuverlässig und besitzt Pflichtbewusstsein	3.37	.67
	ist loyal dem Chef und seinen Unterstellten gegenüber	3.35	.66
	ist verantwortungs- und pflichtbewusst	3.33	.68
	ist ein Vorbild in seiner Erscheinung und seinen Handlungen	3.27	.76
	ist sorgfältig, diszipliniert, zeitgerecht und auftragstreu	3.07	.63
	ist ehrlich sich selbst und anderen gegenüber	3.02	.68
	befolgt, besitzt und vertritt gewisse Werte und Normen	2.97	.76
Belastbarkeit und Engagement	ist psychisch belastbar	3.08	.56
	will eine gute Leistung erzielen	3.07	.61
	ist selbständig und zeigt Eigeninitiative	2.95	.67
	ist von der Sache überzeugt, begeistert; motiviert so die Leute	2.92	.72
	bleibt auch unter Druck sachlich, ausgeglichen und zuverlässig	2.88	.69
	denkt positiv und hat eine positive Grundeinstellung	2.85	.78
Fürsorglichkeit	behandelt seine Unterstellten und die Mitmenschen respektvoll	3.28	.69
	schaut zuerst für seine Leute, stellt sich und seine Bedürfnisse in den Hintergrund	3.08	.79
	kennt seine Unterstellten	3.05	.77
	ist fürsorglich gegenüber seinen Unterstellten	3.00	.71
Konfliktverhalten und Kommunikation	kommuniziert offen, direkt und ehrlich	3.19	.66
	stellt sich dem Konflikt und ist kritikfähig	2.97	.58
	kann mit Konflikten umgehen, erkennt und löst sie	2.93	.61
	kann mit Kritik umgehen, nimmt Korrekturen und Hinweise auf	2.92	.65
Verantwortungsübernahme und Durchsetzungsfähigkeit	übernimmt Verantwortung, besitzt Verantwortungsbewusstsein	3.42	.59
	steht für den Befehl ein	3.19	.78
	trägt die Konsequenzen	3.13	.70
	handelt konsequent und zielorientiert	3.08	.65
	strebt die Zielerreichung an	3.08	.65
	duldet unkorrektes Verhalten nicht	3.03	.74
	ist authentisch	3.02	.75
	greift korrigierend ein, falls sich ein Unterstellter falsch verhält	2.98	.65
	führt und behält den Führungsanspruch	2.98	.72
	stellt sich der Situation, auch wenn sie unangenehm ist	2.88	.61
Planungs- und Organisationsfähigkeit	denkt, plant und handelt vorausschauend	3.15	.71
	lernt aus seinen Fehlern und zieht den Mehrwert daraus	3.10	.63
	setzt Prioritäten (Wichtigkeit und Dringlichkeit)	3.08	.74
	kann die Konsequenzen seines Handelns abschätzen	3.00	.69
	macht eine angepasste Lagebeurteilung	2.98	.82
	kann die Konsequenzen einschätzen und abschätzen	2.95	.57
	sucht und findet eine Lösung	2.90	.73

Anmerkung. *N = 60. Scoring der Wichtigkeitseinstufung: 1 = nicht so wichtig; 2 = wichtig; 3 = sehr wichtig; 4 = unabdingbar.*

Indem ich mit der Critical Incident Technique Anforderungsprofile für Gruppen- und Zugführer und anhand von Urteilen von Subject Matter Experts das Basis-Anforderungsprofil erstellt habe, konnte ich den anfänglichen Mangel an differenzierten Anforderungsprofilen für untere Armee-Kader als Grundlage für die Konstruktion des Leadership-Fragebogens beheben. Tabelle 6.9 zeigt den Vergleich zwischen den Dimensionen der bisher dargestellten Anforderungsprofile.

Die Übersichtsdarstellung in Tabelle 6.9 lässt eine relativ grosse Übereinstimmung der verschiedenen Anforderungsprofile erkennen. Auf Grund der anhand der Critical Incident-Workshops erstellten äusserst umfangreichen Listen mit erfolgsrelevanten Verhaltensweisen sind die beiden daraus gebildeten Anforderungsprofile detaillierter ausgefallen als die anderen. Eine von mir nachträglich durchgeführte Kategorisierung der im Workshop der Arbeitsgruppe Kaderselektion genannten Verhaltensweisen deckt sich mit Ausnahme der Organisations- und Planungsfähigkeit und der Selbstreflexion mit allen Dimensionen des anhand der Critical Incident Technique erstellten Anforderungsprofils für Gruppenführer. Somit könnte der Eindruck entstehen, dass auf die Durchführung der Critical Incident-Workshops hätte verzichtet werden können. Dies mag vielleicht zutreffen, wenn man den Vergleich nur auf die Anzahl und die Benennung der Dimensionen des Anforderungsprofils beschränkt. Der Hauptvorteil des mit der Critical Incident Technique gebildeten Anforderungsprofils liegt jedoch in den die einzelnen Dimensionen charakterisierenden erfolgsrelevanten Verhaltensweisen. Diese leisten bei der Planung und Entwicklung der Selektionsinstrumente sehr gute Dienste, indem sie das zu erfassende Verhalten genau definieren.

In der Arbeitsgruppe Kaderselektion diente das Basis-Anforderungsprofil für unteres Kader als Grundlage für die Erstellung des definitiven Anforderungsprofils. Dabei legten die Experten ein zusätzliches Augenmerk auf die Erfassung der einzelnen Dimensionen anlässlich der Rekrutierung und während der allgemeinen Grundausbildung in der Rekrutenschule. Als Ergebnis dieser Diskussion entstand schlussendlich das in der letzten Spalte der Tabelle 6.9 abgebildete definitive Basis-Anforderungsprofil, welches integriert in eine *Multi-Trait-Multi-Method-Matrix* (Campbell & Fiske, 1959) als Grundlage für den gesamten Selektionsprozess für das untere Milizkader dient und in dieser Form in die entsprechenden Reglemente übernommen wird.

Tabelle 6.9

Vergleich der verschiedenen Anforderungsprofile für unteres Kader der Schweizer Armee

Hoenle / Annen	Arbeitsgruppe Kaderselektion	Gruppenführer (CIT)	Zugführer (CIT)	Basis-Anforderungsprofil	Definitives Basis-Anforderungsprofil
kognitive Leistungsfähigkeit	–	Analysefähigkeit	Analysefähigkeit	–	Auffassungsgabe
–	–	Organisations- & Planungsfähigkeit	Organisations- & Planungsfähigkeit	Planungs- & Organisationsfähigkeit	Organisationsfähigkeit
–	–	–	Allgemeinbildung	–	–
Leistungsmotivation	Belastbarkeit & Leistungsbereitschaft	Leistungsbereitschaft & Engagement	Leistungsbereitschaft & Engagement	Belastbarkeit & Engagement	Engagement & Eigeninitiative
Belastbarkeit		physische & psychische Belastbarkeit	physische & psychische Belastbarkeit		Belastbarkeit & Beharrlichkeit
Gewissenhaftigkeit	–	Gewissenhaftigkeit	Gewissenhaftigkeit & Loyalität	Gewissenhaftigkeit & Loyalität	Gewissenhaftigkeit (Pflichtbewusstsein)
–	–	Offenheit & Flexibilität	Offenheit & Flexibilität	–	–
Durchsetzungsfähigkeit	Durchsetzungsvermögen	Durchsetzungsfähigkeit	Durchsetzungsfähigkeit	Verantwortungsübernahme & Durchsetzungsfähigkeit	Durchsetzungsfähigkeit
Verantwortungsbewusstsein	Verantwortungsbereitschaft	Verantwortungsübernahme	Verantwortungsübernahme		Führungs- und Verantwortungsbereitschaft
–	–	Fürsorglichkeit	Einfühlungsvermögen	Fürsorglichkeit	
–	–	–	Konflikt- & Kritikfähigkeit	Konfliktverhalten & Kommunikation	Konfliktverhalten
Kommunikationsfähigkeit	Teamfähigkeit, Kontakt- & Kommunikationsbereitschaft	Kommunikationsfähigkeit	Kommunikationsfähigkeit		Kommunikation
Kontaktfähigkeit		Teamfähigkeit	Kooperationsfähigkeit	–	Kontaktverhalten
Selbstsicherheit	–	Selbstsicherheit	Auftreten als Chef	–	–
–	–	Selbstreflexion	Selbstreflexion	–	–

Anmerkung. *CIT = Critical Incident Technique. Die Dimensionen des Leadership-Fragebogens sind eingefärbt. Als zusätzliche Dimensionen beim definitiven Anforderungsprofil kommen noch die beiden Aspekte „körperliche Leistungsfähigkeit" und „Interesse an einer militärischen Weiterausbildung" hinzu.*

In Tabelle 6.9 habe ich die Dimensionen markiert, welche auch im Leadership-Fragebogen enthalten sind. Es ist ersichtlich, dass Durchsetzungsfähigkeit und Verantwortungsbewusstsein – welches im Fragebogen auch den Aspekt der Fürsorglichkeit umfasst –, in allen Anforderungsprofilen vorkommen. Auffallend ist jedoch, dass in dem anhand der Einstufungen der Berufsmilitärs gebildeten Basis-Anforderungsprofil – im Gegensatz zu den anderen Anforderungsprofilen – der Aspekt des Zusammenarbeitens (Kontakt-, Kooperations- oder Teamfähigkeit) nicht enthalten ist. In Tabelle 6.10 sind die Verhaltensweisen zur Teamfähigkeit, welche ich den Berufsmilitärs zur Einstufung vorgelegt habe, und deren Wichtigkeitseinstufungen aufgeführt. Die Ränge der einzelnen Verhaltensweisen, welche zwischen 68 und 105 (letzter Rang = 112) liegen, machen deutlich, dass die Berufsmilitärs die Wichtigkeit dieser Aspekte im Vergleich zu den anderen 104 Aspekten als deutlich unterdurchschnittlich eingestuft haben. Absolut betrachtet erreicht keine der acht Verhaltensweisen auch nur annähernd eine durchschnittliche Wichtigkeit von 3, was der Kategorie „sehr wichtig" entspricht.

Tabelle 6.10
Einstufung der Wichtigkeit der Verhaltensweisen zur Teamfähigkeit

Verhaltensweise	Wichtigkeit		
	Rang	*M*	*SD*
ist kontaktfreudig und geht auf Menschen zu	68	2.63	.78
kommuniziert mit seinen Kameraden, um ans Ziel zu kommen	74	2.60	.83
kann sich ein- und unterordnen	75	2.60	.72
vermittelt weiter, wenn er selbst nicht helfen kann	87	2.43	.74
integriert sich in eine Gruppe	88	2.43	.65
kann mit unterschiedlichsten Personen angemessen umgehen	89	2.43	.62
stellt Gruppenkohäsion her und integriert alle ins Team	94	2.37	.64
bietet seine Hilfe an	105	2.25	.68

Anmerkung. *N = 60.*

Über die Gründe, weshalb die Berufsmilitärs zu dieser Einschätzung gelangt sind, lässt sich nur spekulieren – sie müssten in persönlichen Gesprächen abgeklärt werden. Um zu erfahren, ob die „Jobinhaber", also aktive Gruppen- und Zugführer, ebenfalls zu dieser Einschätzung gelangen, habe ich ihnen dieselbe Befragung wie den Berufsmilitärs vorgelegt. Die Ergebnisse aus dieser Zusatzstudie sind auch insofern interessant, als dass es empirische Evidenz gibt, dass sich die Einschätzungen eines kleinen Expertenteams nicht stark von den Einschätzungen einer grossen Gruppe von Jobinhabern unterscheiden (*committee-based vs. field-based job analyses*: z. B. Maurer & Tross, 2000; Tannenbaum & Wesley, 1993).

Im vorliegenden Fall könnte es aber so sein, dass die Berufsmilitärs auf Grund ihrer speziellen Position innerhalb der Armee bezüglich der erwünschten Eigenschaften von unterem Milizkader andere Schwerpunkte setzen als eben diese.

In Tabelle 6.11 ist die Zusammensetzung der Stichprobe der Befragung der Gruppen- und Zugführern dargestellt, welche kurz vor dem Abschluss der Verbandsausbildung in der Rekrutenschule standen, also ihre Ausbildung für die jeweilige Kaderstufe abgeschlossen hatten. Von den in den Lehrverbänden Infanterie, Panzer/Artillerie, Führungsunterstützung Luftwaffe und Fliegerabwehr insgesamt 200 ausgefüllten Fragebogen musste ich knapp einen Viertel ausschliessen, weil diese nicht vollständig oder unseriös ausgefüllt wurden. Auf Grund der vielen von mir im militärischen Umfeld durchgeführten Datenerhebungen weiss ich, dass bei einer befohlenen Teilnahme etliche Rekruten und Kaderangehörige aus Motivationsmangel den Fragebogen unseriös ausfüllen. Dieses Phänomen tritt verstärkt dann auf, wenn die Datenerhebung gegen Ende der Dienstleistungsperiode stattfindet, wie es hier der Fall war. Um einen zusätzlichen Anhaltspunkt für die Beurteilung der Seriosität des Einstufens der Verhaltensweisen zu erhalten, habe ich deshalb die Kontrollverhaltensweise „Kennt die Grundlagen der Einsatzdoktrin auf Stufe Bataillon und Brigade" eingefügt. Vor allem auf der Stufe Gruppenführer ist dieses Wissen nicht von Bedeutung und wird – wenn überhaupt – auch nur am Rande ausgebildet. So schloss ich von den verbleibenden 153 Fragebogen bei den Gruppenführer zusätzlich diejenigen aus, bei welchen die Kontrollverhaltensweise als „sehr wichtig" oder „unabdingbar" eingestuft wurde (n = 41) und bei den Zugführern diejenigen mit der Einstufung „unabdingbar" (n = 6). Somit setzt sich die definitive Stichprobe aus 55 von Gruppenführern (Durchschnittsalter 20.76 Jahre, SD = 1.15) und 51 von Zugführen (Durchschnittsalter 21.12 Jahre, SD = 1.56) ausgefüllten Fragebogen zusammen.

Tabelle 6.11

Beschreibung der Stichprobe der Gruppen- und Zugführer zur Einstufung der Wichtigkeit der Anforderungsdimensionen

Lehrverband	insgesamt ausgefüllt	ungültig ausgefüllt	gültig ausgefüllt		def. Stichprobe*	
			Grfhr	Zfhr	Grfhr	Zfhr
Infanterie	82	14	45	23	21	23
Panzer/Artillerie	66	28	27	11	18	7
Führungsunterstützung Luftwaffe	32	3	18	11	13	11
Fliegerabwehr	20	2	6	12	3	10
Total	200	47	96	57	55	51

Anmerkung. *Grfhr = Gruppenführer, Zfhr = Zugführer. *Ausschluss auf Grund der Beantwortung der Kontrollverhaltensweise: Grfhr = sehr wichtig, unabdingbar; Zfhr = unabdingbar.*

Tabelle 6.12

Von den Gruppenführern eingestufte Wichtigkeit der Anforderungsdimensionen

Dimension	*Range*	*M*	*SD*	*M (1/8)*	*M Rang*	*SD*
Auftreten als Chef / Selbstsicherheit	10 – 29	22.55	4.18	2.82	37.50	36.23
Verantwortungsübernahme	10 – 31	22.49	3.94	2.81	35.88	33.88
Leistungsbereitschaft & Engagement	10 – 31	21.67	4.36	2.71	45.25	27.04
Planungs- & Organisationsfähigkeit	12 – 27	21.64	2.64	2.71	44.88	22.67
Gewissenhaftigkeit & Loyalität	13 – 27	21.58	3.20	2.70	50.38	42.56
physische & psychische Belastbarkeit	10 – 28	21.51	3.44	2.69	48.13	24.71
Durchsetzungsfähigkeit	10 – 30	21.40	3.65	2.68	47.38	34.70
Teamfähigkeit	9 – 30	21.13	3.93	2.64	56.38	21.77
Fürsorglichkeit / Einfühlungsvermögen	9 – 32	21.00	4.27	2.63	55.88	38.10
Kommunikationsfähigkeit	8 – 27	20.64	3.79	2.58	62.00	32.07
Analysefähigkeit	12 – 27	20.25	2.81	2.53	72.00	23.83
Konflikt- & Kritikfähigkeit	9 – 26	19.87	3.52	2.48	70.38	34.76
Selbstreflexion	8 – 28	19.67	3.67	2.46	78.13	28.90
Offenheit & Flexibilität	11 – 26	19.36	3.03	2.42	86.88	17.08

Anmerkung. *N = 55. Scoring der Wichtigkeitseinstufung:*
1 = nicht so wichtig; 2 = wichtig; 3 = sehr wichtig; 4 = unabdingbar.

Tabelle 6.13

Von den Zugführern eingestufte Wichtigkeit der Anforderungsdimensionen

Dimension	*Range*	*M*	*SD*	*M (1/8)*	*M Rang*	*SD*
Verantwortungsübernahme	18 – 32	25.22	3.32	3.15	25.75	28.00
Auftreten als Chef / Selbstsicherheit	15 – 30	24.57	3.48	3.07	36.63	30.05
Planungs- & Organisationsfähigkeit	17 – 32	24.53	3.18	3.07	31.13	23.34
Leistungsbereitschaft & Engagement	14 – 32	24.20	4.39	3.03	35.50	30.53
Gewissenhaftigkeit & Loyalität	17 – 32	23.88	3.39	2.99	45.13	34.24
Durchsetzungsfähigkeit	13 – 31	23.75	3.51	2.97	43.88	23.41
physische & psychische Belastbarkeit	17 – 32	23.27	3.29	2.91	54.00	30.46
Fürsorglichkeit / Einfühlungsvermögen	16 – 32	22.94	4.02	2.87	63.38	32.57
Analysefähigkeit	16 – 32	22.94	4.02	2.87	62.38	16.41
Kommunikationsfähigkeit	14 – 31	22.55	3.95	2.82	68.63	19.86
Selbstreflexion	14 – 30	22.04	4.10	2.76	77.25	33.89
Offenheit & Flexibilität	13 – 29	21.92	4.40	2.74	78.88	26.73
Teamfähigkeit	12 – 29	21.53	4.16	2.69	81.88	25.45
Konflikt- & Kritikfähigkeit	14 – 32	21.33	4.47	2.67	86.63	21.65

Anmerkung. *N = 51. Scoring der Wichtigkeitseinstufung:*
1 = nicht so wichtig; 2 = wichtig; 3 = sehr wichtig; 4 = unabdingbar.

Die von den Gruppen- und Zugführern für ihre jeweilige Kaderstufe vorgenommenen Einstufungen der Wichtigkeit der 14 Anforderungsdimensionen habe ich in den Tabellen 6.12 und 6.13 dargestellt. (Die Einstufungen zu allen 112 Verhaltensweisen sind in den Anhängen 6.8 bis 6.11 aufgeführt.) Wie bei den Berufsmilitärs fallen auch hier die anhand der Intraklassenkorrelation berechneten Beurteilerreliabilitäten sehr tief aus: ICC_{unjust} = .09 (F(111, 5'550) = 6.93, p < .001) respektive ICC_{unjust} = .09 (F(111, 5'217) = 7.08, p < .001), die über alle Beurteiler gemittelten Reliabilitäten sind jedoch mit $ICC_{unjust,MW}$ = .83 respektive $ICC_{unjust,MW}$ = .82 als gut zu bezeichnen (Wirtz & Caspar, 2002). Damit fallen die Reliabilitäten bei den „Jobinhaber" deutlich tiefer aus als diejenigen bei den Berufsmilitärs ($ICC_{unjust,MW}$ = .91), was sich jedoch mit dem Ergebnis aus der Meta-Analyse von Dierdorff und Wilson (2003) deckt. Zur besseren Vergleichbarkeit habe ich die Rangfolgen der Einstufungen der Gruppen- und Zugführer und der Berufsmilitärs in Tabelle 6.14a einander gegenübergestellt, wobei ich Unterschiede in den Einstufungen von drei oder mehr Rangplätzen farblich hervorgehoben habe. Dabei fällt die unterschiedliche Rangierung der Anforderungsdimension Teamfähigkeit auf: Wie die Berufsmilitärs haben die Zugführer diese im Vergleich zu den anderen Anforderungsdimensionen als unwichtig eingestuft, die Gruppenführer jedoch im Mittelfeld. Die Dimensionen Leistungsbereitschaft & Engagement, physische & psychische Belastbarkeit und Teamfähigkeit haben die Gruppen- und Zugführer als wichtiger eingestuft als die Berufsmilitärs, welche ihrerseits Gewissenhaftigkeit & Loyalität und Fürsorglichkeit / Einfühlungsvermögen als wichtiger erachten.

Um die Unterschiede der Einstufungen der Anforderungsdimensionen zwischen den drei Gruppen auf statistische Signifikanz hin untersuchen zu können, habe ich diese auf Personenebene rangiert. Dieser Schritt ist notwendig, da sich die durchschnittlichen Einstufungen in den drei Gruppen unterscheiden: Sie liegt bei den Gruppenführern bei M = 2.63 (SD = .34), bei den Zugführern bei M = 2.90 (SD = .41) und bei den Berufsmilitärs bei M = 2.71 (SD = .34). Die auf diese Weise bestimmten Reihenfolgen habe ich in Tabelle 6.14b abgebildet. Der Vergleich der Rangreihenfolgen der Gruppen- mit denjenigen der Zugführern ergibt lediglich bei der Anforderungsdimension Teamfähigkeit einen signifikanten Unterschied (U = 797.50, z = -3.83, p < .001, r = -.37). Beim Vergleich der Rangreihenfolgen der Gruppen- und Zugführern mit denjenigen der Berufsmilitärs zeigen sich in fünf Dimensionen signifikante Unterschiede: Bei der physischen & psychischen Belastbarkeit (U = 1'937.00, z = -4.18, p < .001, r = -.32), der Teamfähigkeit (U = 2'159.50, z = -3.44, p < .01, r = -.27), der Gewissenhaftigkeit & Loyalität (U = 2'205.50, z = -3.28, p < .01, r = -.25) und der Fürsorglichkeit / Einfühlungsvermögen (U = 2'460.50, z = -2.42, p < .05, r = -.19).

Tabelle 6.14a

Vergleich der rangierten Skalenmittelwerte der von den verschiedenen Kadergruppen eingestuften Wichtigkeit der einzelnen Anforderungsdimensionen

	Gruppen- & Zugführer (*n* = 106)	Berufsmilitärs (*n* = 60)	Gruppenführer (*n* = 55)	Zugführer (*n* = 51)
Verantwortungsübernahme	1	1	2	1
Auftreten als Chef / Selbstsicherh.	2	3	1	2
Planungs- & Organisationsfähigkeit	3	4	4	3
Leistungsbereitschaft & Engagem.	4	7	3	4
Gewissenhaftigkeit & Loyalität	5	2	5	5
Durchsetzungsfähigkeit	6	6	7	6
physische & psych. Belastbarkeit	7	11	6	7
Fürsorglichkeit / Einfühlungsverm.	8	5	9	8
Kommunikationsfähigkeit	9	9	10	10
Analysefähigkeit	10	8	11	9
Teamfähigkeit	11	14	8	13
Selbstreflexion	12	10	13	11
Offenheit & Flexibilität	13	12	14	12
Konflikt- & Kritikfähigkeit	14	13	12	14

Anmerkung. Unterschiede von drei oder mehr Rangplätzen sind mit Farbe hinterlegt.

Tabelle 6.14b

Vergleich der Rangreihenfolgen der von den verschiedenen Kadergruppen eingestuften Wichtigkeit der einzelnen Anforderungsdimensionen

	Gruppen- & Zugführer (*n* = 106)	Berufsmilitärs (*n* = 60)	Gruppenführer (*n* = 55)	Zugführer (*n* = 51)
Verantwortungsübernahme	1	1	1	1
Auftreten als Chef / Selbstsicherh.	2	3	2	3
Planungs- & Organisationsfähigkeit	3	4	3	2
Leistungsbereitschaft & Engagem.	4	6	7	4
Gewissenhaftigkeit & Loyalität	5	2	4	5
Durchsetzungsfähigkeit	6	7	5	6
physische & psych. Belastbarkeit	7	11	6	7
Fürsorglichkeit / Einfühlungsverm.	8	5	9	8
Kommunikationsfähigkeit	9	8	10	10
Analysefähigkeit	10	9	11	9
Teamfähigkeit	11	14	8	13
Selbstreflexion	12	10	13	11
Konflikt- & Kritikfähigkeit	13	12	12	14
Offenheit & Flexibilität	14	13	14	12

Anmerkung. Signifikante Unterschiede in den Rangreihenfolgen sind mit Farbe hinterlegt.

Deutliche und signifikante Mittelwertsunterschiede (U = 2'183.50, z = -3.37, p < .01, r = -.26) zeigen sich bei der Verantwortungsübernahme, obwohl alle drei Gruppen diese als wichtigste Anforderungsdimension eingestuft haben: Bei den Gruppenführern liegt deren durchschnittlicher Rangplatz bei M = 5.30 (SD = 3.44), bei den Zugführern bei M = 4.21 (SD = 2.97), bei den Berufsmilitärs bei M = 3.27 (SD = 2.77) und bei den Gruppen- und Zugführern zusammen bei M = 4.77, (SD = 3.25). Nicht signifikant wird hingegen der Unterschied in der Anforderungsdimension Leistungsbereitschaft & Engagement (U = 2'884.00, z = -1.00, p = .32, r = -.08).

Tabelle 6.15

Unterschiede in der von den Gruppen- und Zugführern der verschiedenen Lehrverbänden eingestuften Wichtigkeit der einzelnen Anforderungsdimensionen

	Gruppenführer			Zugführer		
	Inf *n* = 21	Pz/Art *n* = 18	FULW *n* = 13	Inf *n* = 23	Pz/Art *n* = 7	FULW *n* = 11
Verantwortungsübernahme	2	1	7	1	2	1
Auftreten als Chef / Selbstsicherh.	1	6	2	2	3	2
Planungs- & Organisationsfähigkeit	7	4	3	4	1	3
Leistungsbereitschaft & Engagem.	4	5	10	3	5	4
Gewissenhaftigkeit & Loyalität	5	2	9	6	4	6
Durchsetzungsfähigkeit	3	3	14	5	7	5
physische & psych. Belastbarkeit	8	7	5	8	6	7
Fürsorglichkeit / Einfühlungsverm.	10	11	1	7	9	9
Kommunikationsfähigkeit	9	10	8	10	11	8
Analysefähigkeit	12	8	11	9	8	11
Teamfähigkeit	6	9	4	12	13	14
Selbstreflexion	11	13	12	11	10	12
Offenheit & Flexibilität	13	14	13	13	14	10
Konflikt- & Kritikfähigkeit	14	12	6	14	12	13
Durchschnittliche Wichtigkeit der Verhaltensweisen (*M* / *SD*)	2.74 .24	2.60 .40	2.52 .36	2.98 .37	2.59 .50	2.81 .27

Anmerkung. *Inf = Lehrverband Infanterie; Pz/Art = Lehrverband Panzer/Artillerie; FULW = Lehrverband Führungsunterstützung Luftwaffe. Die Teilstichprobe aus dem Lehrverband Fliegerabwehr habe ich auf Grund des geringen Umfangs bei den Gruppenführern (n = 3) nicht aufgeführt. Signifikante Unterschiede in den Rangreihenfolgen sind mit Farbe hinterlegt.*

Wie in Tabelle 6.15 ersichtlich ist, sind die Unterschiede zwischen den vier Truppengattungen respektive Lehrverbänden, aus welcher sich die Stichprobe mit den Gruppen- und Zugführern zusammensetzt, viel grösser als diejenigen zwischen

den drei Kader-Gruppen. So stufen zum Beispiel die Gruppenführer der Lehrverbände Infanterie und Panzer/Artillerie die Verantwortungsübernahme als eines der wichtigsten Kriterien ein, für die Gruppenführer des Lehrverbandes Führungsunterstützung Luftwaffe belegt dieses Kriterium den Rang sieben. Umgekehrt ist für diese Fürsorglichkeit sehr wichtig, ein Kriterium, welches bei den beiden anderen Gruppen die Rangplätze zehn und elf belegt. Signifikante Unterschiede in den Rangreihenfolgen der Gruppenführer der drei Lehrverbände ergeben sich bei den Anforderungsdimensionen Verantwortungsübernahme ($H(2) = 8.67$, $p < .05$), Durchsetzungsfähigkeit ($H(2) = 14.44$, $p < .01$) und Fürsorglichkeit / Einfühlungsvermögen ($H(2) = 13.11$, $p < .01$). Bei den Zugführern unterscheiden sich die drei Rangierungen nicht signifikant voneinander.

Maurer und Tross (2000) führen als eine Gefahr bei einem alleinigen Einsatz von kleinen Expertenteams zur Anforderungsanalyse auf, dass diese eventuell nicht alle Aspekte des zu beurteilenden Arbeitsgebietes abdecken. Lievens, Sanchez und De Corte (2004) schlagen auf Grund der Ergebnisse ihrer Studie vor, bei der Anforderungsanalyse *Subject Matter Experts* einzubinden, welche unterschiedliche Ansichten oder Kontakte mit der einzustufenden Arbeitsstelle haben, wie zum Beispiel Jobinhaber, Vorgesetzte oder interne Kunden. Auch bei den von mir erhobenen Daten unterscheidet sich das von den Berufsmilitärs abgegebene Expertenrating bei knapp einem Drittel der Anforderungsdimensionen von demjenigen der Funktionsinhaber. Hingegen unterscheiden sich die Einstufungen durch die Gruppenführer kaum von denjenigen durch die Zugführer. Auffallend gross sind jedoch die Unterschiede der Einstufungen der Kader der verschiedenen Lehrverbände: Bei den Gruppenführern beträgt der durchschnittliche Unterschied der Rangplätze der einzelnen Anforderungsdimensionen 3.52 Ränge, bei den Zugführern hingegen nur 1.48. Zudem unterscheiden sich – wie oben dargestellt – die Einstufungen der Gruppenführer der drei Lehrverbände in drei Anforderungsdimensionen signifikant voneinander, bei den Zugführern ergeben sich jedoch keine Unterschiede. Dies könnte darauf zurückzuführen sein, dass sich die Tätigkeit der Gruppenführer nicht auf das reine Befehlen der anvertrauten Soldaten beschränkt, sondern dass sie bei der Arbeit auch mitanpacken müssen und somit Unterschiede zwischen den Tätigkeiten der verschiedenen Truppengattungen bei der Einstufung der Wichtigkeit der führungsbezogenen Anforderungsdimensionen eine bedeutende Rolle spielen. Mit zunehmender Ranghöhe nehmen jedoch die truppenspezifischen Unterschiede ab und übergeordnete Führungsaufgaben treten in den Vordergrund.

Eine alternative Erklärung für unterschiedliche Einstufungen bei gleichen oder ähnlichen Tätigkeitsgebieten sind Unterschiede in der Organisationskultur: Li, Wang, Taylor, Shi und He (2008) konnten anhand ihrer Studie mit 270 Kundenberatern von 37 Mobiltelefongesellschaften aufzeigen, dass sich die Einstufungen der arbeitsrelevanten Persönlichkeitsdimensionen Leistungsmotivation und Gewissenhaftigkeit zwischen den Gesellschaften unterscheiden und in einem Zusammenhang mit zwei korrespondierenden Dimensionen der Organisationskultur stehen. Auch die Kommandanten der Rekrutenschulen wirken im Rahmen ihrer Möglichkeiten prägend auf die Organisationskultur, so dass sich diesbezüglich Unterschiede zwischen den verschiedenen Schulen herausbilden. Dass sich die Einstufungen der Zugführer der drei untersuchten Lehrverbände jedoch nicht signifikant voneinander unterscheiden, ist ein Hinweis darauf, dass die Organisationskulturen vergleichbar sind und die gefundenen Unterschiede bei den Gruppenführern eher durch die unterschiedlichen Tätigkeiten bedingt sind. Dierdorff und Morgeson (2007; siehe auch Lievens, Sanchez, Bartram & Brown, 2010) untersuchten Aspekte des organisationalen Kontextes als mögliche Ursachen für Unterschiede in der Einstufung von Arbeitsanforderungen und konnten aufzeigen, dass geringe Autonomie, hohe Abhängigkeit und ein sich stark wiederholender Arbeitsablauf zu einem höheren Konsens bei der Verhaltenseinstufung führen. Diese Faktoren treffen grösstenteils auch auf die Gruppenführer zu, weshalb auch sie keine Erklärung für die gefundenen Unterschiede liefern.

Dass sich die von verschiedenen Personen – seien es Experten oder Jobinhaber – vorgenommenen Einstufungen von Anforderungen einer Arbeitsstelle deutlich voneinander unterscheiden, ist ein allgemein auftretendes Phänomen. Ursprünglich wurde angenommen, dass es sich dabei lediglich um Fehlervarianz handelt, welche sich bei der Verwendung genügend grosser Stichproben aufhebt (z. B. Green & Stutzman, 1986; Sanchez & Frazer, 1992). In aktuellen Artikeln verstehen die Forscher diese Unterschiede jedoch als Ausdruck davon, dass die Jobinhaber einerseits ihre Arbeitsstelle verschieden wahrnehmen und andererseits bewusst oder unbewusst ein mit den persönlichen Motiven kongruentes Bild ihrer Arbeitsstelle abgeben (z. B. Lievens et al., 2010; Morgeson & Campion, 1997). So nennen Dierdorff und Wilson (2003) als einen der Gründe für unterschiedliche Einstufungen die Komplexität der zu beurteilenden Verhaltensdimensionen, welche besonders bei Führungs- und Managementaufgaben hoch ausgeprägt ist. Dabei hängt – wie nicht anders zu erwarten ist – die Schwierigkeit der Einstufung einer Eigenschaft mit der Beurteilerreliabilität zusammen (Wohlers & London, 1989). Dementsprechend sind zum Beispiel Eigenschaften dann schwierig einzuschätzen, wenn sie selten auftreten oder wenn deren Verknüpfung mit konkretem Verhalten nicht evident ist.

Ein ebenfalls gut nachvollziehbarer Befund sind tiefere Interrater-Reliabilitäten bei mit der Anforderungsanalyse-Methode unerfahrenen Ratern (Cornelius, DeNisi & Blencoe, 1984; Lievens et al., 2004; Voskuijl & van Sliedregt, 2002), wobei in der Studie von Lievens und Sanchez (2007) ein Rater-Training (*frame-of-reference training*; Bernardin & Buckley, 1981; siehe auch Schleicher, Day, Mayes & Riggio, 2002) zu einer deutlich verbesserten Akkuratheit der Anforderungsanalyse respektive des *competency modelings* führte. In der Meta-Analyse von Voskuijl und van Sliedregt (2002) zeigte sich jedoch, dass Trainings nur bei Experten die Interrater-Reliabilität erhöhten, nicht jedoch bei Studenten oder Jobinhabern. Keinen Einfluss auf die Reliabilität der Einstufungen hat die Leistungsfähigkeit der Jobinhaber (Conley & Sackett, 1987; Wexley & Silverman, 1978).

Auch das *job crafting* – die vom Jobinhaber vorgenommenen Änderungen im Aufgabenspektrum ihres Arbeitsgebietes (Wrzesniewski & Dutton, 2001) – kann zu unterschiedlichen Einstufungen der Anforderungen einer Berufstätigkeit führen. Dieser Effekt lässt sich in dieser Studie hingegen ausschliessen, da vor allem bei den Gruppenführern kaum individueller Spielraum bei den auszuführenden Tätigkeiten besteht.

Morgeson und Campion (1997) führen insgesamt 16 soziale und kognitive Ursachen auf, welche zu Verzerrungen bei der Befragung von Jobinhabern führen können, was sich in einer ungenauen Arbeitsanalyse auswirken kann. Zwei davon haben auch bei der hier beschriebenen Erhebung eine grosse Rolle gespielt:

Fehlendes Interesse und fehlende Motivation (motivation loss) beim Bearbeiten des Arbeitsanalyse-Fragebogens tritt vor allem dann auf, wenn eine grosse Gruppe befragt wird, sich der Beitrag des Einzelnen nicht überprüfen lässt und/oder die Teilnehmer den Sinn der Befragung nicht erkennen. Dabei treten die aus den Studien im Zusammenhang mit der Leistung von Gruppen bekannten Phänomene auf: Die als Trittbrettfahren (*free-riding effect*) bezeichnete Verringerung der Leistung auf Grund der Annahme, dass der eigene Beitrag nur einen kleinen Effekt auf das Gruppenergebnis hat (z. B. Albanese & Van Fleet, 1985) und das soziale Faulenzen (*social loafing*), die Verringerung der Leistung auf Grund der Tatsache, dass der eigene Beitrag nicht identifizierbar ist (z. B. Kidwell & Bennett, 1993). Weldon und Gargano (1985) konnten den Effekt der Verantwortungsdiffusion im Zusammenhang mit der Evaluation von Arbeitsstellenbeschreibungen nachweisen, indem sich in ihrer Studie zeigte, dass Probanden, welche instruiert wurden, dass nur sie die Einstufung vornehmen, beim Bearbeiten der entsprechenden Fragebogen mehr leisteten und auch einen höheren kognitiven Aufwand betrieben als solche, welche sich als Teil einer Gruppe von

Evaluatoren glaubten. Diese Phänomene haben mit Sicherheit auch bei meiner Befragung eine Rolle gespielt, da die Gruppen- und Zugführer die Fragebogen in Gruppen ausgefüllt haben und es so für jeden Teilnehmer klar war, dass sein Beitrag nur zu einem kleinen Teil die Studienergebnisse direkt beeinflusst. Zudem führte ich die Befragung anonym durch, so dass also der Rater eines offensichtlich unseriös ausgefüllten Fragebogens (z. B. immer dieselbe Antwortalternative gewählt oder mit den Antwortkreuzen produzierte Muster) sicher sein konnte, dass es unmöglich ist, ihn zu eruieren. Der Umstand, dass die Befragung in der letzten Woche der Rekrutenschule durchgeführt wurde, trägt weiter dazu bei, dass die Motivation, konzentriert und seriös eine Befragung auszufüllen, bei einigen Gruppen- und Zugführern sehr tief war.

Unseriöses Ausfüllen (carelessness) kann als Folge der oben beschriebenen fehlenden Motivation auftreten, kann aber durchaus auch andere Ursachen haben, wie zum Beispiel sehr ausführliche Analysefragebogen, deren Bearbeitung bis zu zwei Stunden in Anspruch nehmen kann, komplexe oder unpassende Items oder ein Unverständnis seitens des Teilnehmers darüber, was von ihm verlangt wird. In der Literatur werden mehrere Verfahren beschrieben, um unseriös vorgenommene Einstufungen in Arbeitsanalyse-Fragebogen aufzuspüren: So zum Beispiel die Wiederholung einzelner Items im Fragebogen (Wilson, Harvey & Macy, 1990), das Berechnen der Übereinstimmung der Einstufungen eines Raters mit den anderen Ratern (Hughes & Prien, 1989), das Einfügen von für die untersuchte Arbeitsstelle unwichtiger Verhaltensweisen (Green & Stutzman, 1986) oder ein – mit der PRF-Infrequenz-Skala oder der MMPI-Skala F vergleichbarer – Infrequenz-Index (Green & Veres, 1990). Dabei zeigte sich, dass der Ausschluss der Fragebogen unseriös ausfüllender Rater zu einer deutlichen Erhöhung der Reliabilität der Messung führt (Green & Stutzman, 1986; Wilson et al., 1990). Das Problem mit unseriös ausgefüllten Fragebogen ist mir im Zusammenhang mit Datenerhebungen in Rekrutenschulen bestens bekannt und stellt eine grosse Herausforderung beim Datenscreening dar. In dieser Studie habe ich 23.50% der Fragebogen von der weiteren Verarbeitung ausgeschlossen, weil die Studienteilnehmer diese offensichtlich unseriös ausgefüllt hatten. Zudem habe ich 42.71% der verbleibenden Fragebogen von Gruppen- und 10.53% derjenigen von Zugführern nicht in die Berechnungen einbezogen, weil sie die Wichtigkeit der Kontrollverhaltensweise viel zu hoch eingestuft haben. Insgesamt habe ich nur gerade 53.00% der ursprünglich ausgefüllten Fragebogen in den Auswertungen berücksichtigt. Eine ähnlich hohe Ausschlussquote ergab sich auch in der Studie von Green und Stutzman (1986), in welcher 57% der Studienteilnehmer angegeben haben, dass sie Zeit für die Ausführung einer Aufgabe aufwendeten, welche sie unmöglich ausgeführt haben können.

Alle in diesem Kapitel aufgeführten Recherchen und Datenerhebungen belegen, dass die drei Verhaltensdimensionen Durchsetzungsfähigkeit, Kontaktfähigkeit und Verantwortungsbewusstsein wichtige – wenn zum Teil auch nicht die wichtigsten – Aspekte der Führungskompetenz von unterem respektive unterstem Milizkader der Schweizer Armee abdecken und es somit gerechtfertigt ist, diese in der Kaderselektion zu erheben und zu beurteilen.

Dieser hier beschriebene Arbeitsschritt ist bei nach der etablierten Methode entwickelten Situational Judgment Tests nicht vorgesehen. An seiner Stelle erheben die Testentwickler erfolgskritische Verhaltensweisen, welche ihnen als Gesamtheit und ohne weitere Unterteilung in Verhaltensdimensionen als Ausgangsmaterial für die Bildung der Item-Stämme dienen. Für die Entwicklung des Leadership-Fragebogens sammelten wir nun nicht Verhaltensweisen, welche generell indikativ für erfolgreiche Kader sind, sondern liessen Probanden Verhaltensweisen zu den drei ausgewählten Dimensionen schildern. Dazu setzen wir, wie im nachfolgenden Kapitel ausführlich beschrieben, den *Act Frequency Approach* ein.

6.2 Generierung der Item-Stämme

Klassische Persönlichkeits-Fragebogen unterscheiden sich von einem situativen Test dahingehend, als dass es dem Bearbeiter überlassen wird, sein Verhalten transsituational zu beschreiben, indem er sich verschiedene in der Vergangenheit konkret erlebte Situationen vorstellt und sich Gedanken dazu macht, wie er sich damals verhalten hat. Er muss dann schlussendlich die einzelnen vergegenwärtigten Verhaltensweisen zu einer Gesamteinschätzung aggregieren, zum Beispiel zur Einschätzung „Ich suche aktiv den Kontakt zu meinen Mitmenschen.". Die Testautoren lassen es dem Bearbeiter jedoch offen, welche Situationen er zur Einschätzung der jeweiligen Aussage im Fragebogen heranziehen soll. Setzt der Diagnostiker den Fragebogen nun in einem spezifischen Kontext ein, zum Beispiel in der Personalselektion, so muss er dem Bewerber die Anweisung geben, dass dieser sich bei der Einstufung der Aussagen im Testverfahren Situationen im beruflichen Kontext vorstellen soll. Darauf verweisen die Testautoren in ihren Manualen zum Teil explizit, wie das Beispiel aus den Anweisungen zu den Instruktionen an den Testkandidaten im Manual des Bochumer Inventars zur berufsbezogenen Persönlichkeitsbeschreibung zeigt: „[Es] ist darauf hinzuweisen, dass sich alle Aussagen auf Verhalten und Erleben im Berufsleben beziehen – es

sei denn, mit einer Frage wird ausdrücklich ein anderes, zum Beispiel das Freizeitverhalten, thematisiert" (Hossiep & Paschen, 2003, S. 51).

Bei einem situativen Persönlichkeits-Fragebogen hingegen gibt man dem Bearbeiter konkrete Situationen und eine Auswahl an möglichen Verhaltensweisen in diesen Situationen vor. Dieser muss sich dann nur noch vergleichbare, selbst erlebte Situationen in Erinnerung rufen und sich vergegenwärtigen, wie er sich damals verhalten hat. Gelingen kann dies jedoch nur, wenn die gewählten Situationen auch mit einer hohen Wahrscheinlichkeit dem Erfahrungsschatz der Bearbeiter entsprechen. Somit ist es beinahe zwangsläufig so, dass situative Persönlichkeits-Fragebogen massgeschneidert für den späteren Einsatzzweck und die intendierte Testpopulation konstruiert werden müssen. Es macht tatsächlich wenig Sinn, „ ... 20jährigen Lehrabsolventen denselben Test vorzulegen wie Führungskräften, welche sich auf eine Top-Kaderstelle bewerben. Vor allem situative Verfahren müssen auf den Erfahrungsschatz der Bewerber abgestimmt sein" (Boss, 2005, S. 33–34).

Bei der Testung junger Berufseinsteiger steht der Diagnostiker jedoch vor dem Problem, dass diese nur über einen sehr eingeschränkten arbeitsbezogenen Erlebnisschatz verfügen, da sie viele arbeitsrelevante Situationen noch nicht persönlich erlebt haben. Noch deutlicher zeigt sich diese Problematik bei der vordienstlichen Kaderbeurteilung anlässlich der Rekrutierung: Die Stellungspflichtigen kennen den Militärdienst höchstens vom Hörensagen und die meisten von ihnen haben zudem noch keine Führungserfahrungen machen können. Wenn diese nun in einem situativen Persönlichkeitstest ihr Verhalten in erfolgskritischen militärischen Führungssituationen zu beschreiben haben, können sie nicht von in der Vergangenheit in konkreten Situationen gezeigtem Verhalten berichten, sondern sie müssten hypothetische Aussagen dazu machen, wie sie sich in diesen Situationen wohl am ehesten verhalten würden. Dieses Vorgehen entspricht dem Konzept des situativen Interviews (Latham, Saari, Pursell & Campion, 1980). Nun belegen jedoch Forschungsergebnisse– zumindest für anspruchsvollere berufliche Positionen –, dass Verhaltensbeschreibungs-Interviews (Janz, 1982), in welchen die Bewerber Angaben zu in der Vergangenheit gezeigtem Verhalten machen müssen, höhere Validitäten erzielen als situative Interviews (Huffcutt, Conway, Roth & Klehe, 2004; Huffcutt, Weekley, Wiesner, Degroot & Jones, 2001; Pulakos & Schmitt, 1995; Taylor & Small, 2002). Aber auch ganz allgemein gilt in der Eignungsdiagnostik, dass in der Vergangenheit in ähnlichen beruflichen Situationen gezeigtes Verhalten der bester Prädiktor für zukünftiges Verhalten darstellt (z. B. Ajzen, 2002; Janz, 1989; Motowidlo, 1999; Motowidlo et al., 1992; Ouellette & Wood, 1998).

Gegen die Verwendung von erfolgskritischen militärischen Führungssituationen als Grundlage für die Entwicklung eines SJT für den Einsatz in der Rekrutierung der Schweizer Armee spricht zudem, dass ein grosser Anteil der Stellungspflichtigen dem Militärdienst mit gemischten Gefühlen entgegensieht und die Schilderung militärischer Situationen in einem Testverfahren bei einigen zu Reaktanz und somit zu Unruhe im Testsaal führen könnte. Somit muss bei dieser Testentwicklung eine Übertragung von erfolgskritischen militärischen Führungssituationen in die Erlebnisrealität der Stellungspflichtigen erfolgen. Und weil diese in der Regel noch nie eine Führungsposition inne gehabt haben, ist es nicht zielführend, für die militärischen Situationen die jeweiligen Entsprechungen im Zivilleben zu suchen. Aus diesem Grund baut der Leadership-Fragebogen nicht auf erfolgskritischen Führungssituationen auf, sondern – wie in Kapitel 6.1 dargestellt – auf bedeutsamen, in den drei Anforderungsdimensionen zusammengefassten Persönlichkeitseigenschaften für unteres Kader der Schweizer Armee. Zu diesen müssen wir nun Situationen aus dem Alltagserleben von Jugendlichen finden, in welchen diese Persönlichkeitseigenschaften zum Tragen kommen. Um dabei die Erfahrungsrealität der Stellungspflichtigen möglichst genau abzubilden, wählten die mit der Generierung des Ausgangsmaterials beauftragte studentische Arbeitsgruppe und ich den Act Frequency Approach nach Buss und Craik (1980, 1983; ausführlich in Kapitel 3 dargestellt), wobei wir uns stark an das von Krüger und Amelang (1995) bei der Konstruktion ihres Fragebogens zur Erfassung der Risikobereitschaft beschriebene Vorgehen anlehnten. Ziel war es, zu jeder der drei Dimensionen des Leadership-Fragebogens – Durchsetzungsfähigkeit, Kontaktfähigkeit und Verantwortungsbewusstsein – anhand der Angaben Jugendlicher eine Liste von hochprototypischen Verhaltensweisen zu erhalten. Wie Krüger und Amelang beim Konstrukt Risikobereitschaft hatten auch wir bei der Dimension Verantwortungsbewusstsein das Problem einer fehlenden theoretischen Untermauerung, welche uns als Grundlage für die Testentwicklung hätte dienen können. Mit dem Einsatz des Act Frequency Approach konnten wir dieses Problem jedoch umgehen, indem wir uns auf das allgemeine, intuitive Verständnis Jugendlicher zu dieser und den zwei anderen Dimensionen stützten.

Nachfolgend ist unser Vorgehen bei der Generierung der Item-Stämme des Leadership-Fragebogens mittels des Act Frequency Approach (AFA) dargestellt (Deiss, Emerson, Imper & Maier, 2002).

1. Generierung der Acts

Als Grundlage für die Formulierung der Item-Stämme der situativen Aufgaben des Leadership-Fragebogens, sammelte eine studentische Arbeitsgruppe bei ins-

gesamt 38 Schülerinnen und Schüler einer Gymnasial- und einer Berufsschulklasse im Alter von 18 bis 20 Jahren während einer regulären Schulstunde schriftlich Acts zu den drei Persönlichkeitsdimensionen. Die Schülerinnen und Schüler erhielten mündlich und schriftlich folgende Instruktion (hier am Beispiel der Dimension Kontaktfähigkeit):

1. Lesen Sie die Definition „Kontaktfähigkeit" genau durch!
2. Stellen Sie sich je eine weibliche und eine männliche Person aus Ihrem Freundes- oder Bekanntenkreis vor, die Sie als sehr kontaktfähig einschätzen!
3. Überlegen Sie sich für jede dieser Personen je zwei einzelne Handlungen in konkreten Situationen, in denen Ihrer Meinung nach das kontaktfähige Verhalten dieser Person zum Ausdruck gekommen ist!
4. Notieren Sie die vier Handlungen in der Tabelle auf Seite 2!

   Wichtig! Achten Sie bitte darauf, dass Sie beobachtbare Handlungen in konkreten Situationen beschreiben und nicht allgemeine Eigenschaften notieren!

Es folgen je zwei Beispiele für richtig formulierte Acts, also beobachtbare Handlungen in konkreten Situationen, und falsch formulierte Acts (allgemeine Eigenschaftsbegriffe) zu dominantem Verhalten und die jeweilige Konstruktdefinitionen:

Richtig: Er verteilte die Rollen für ein Theaterstück.
Sie verbot einem Freund, den Raum zu verlassen.

Falsch: Er verhält sich dominant.
Sie ist herrschsüchtig.

Definition „*Durchsetzungsfähigkeit*"

Durchsetzungsfähige Menschen können ihre eigenen Interessen gegenüber anderen wahren. Sie sind in der Lage, Widerstände zu überwinden, welche durch andere Personen verursacht sind. Sie vertreten den eigenen Standpunkt mit Nachdruck und überzeugen, indem sie diesen anderen vermitteln.

Definition „*Kontaktfähigkeit*"

Kontaktfähige Menschen ziehen es vor, in Gesellschaft zu sein. Sie suchen die Nähe der Menschen und fühlen sich im grossen Kreise wohl. Sie sind umgänglich, schätzen den geselligen Umgang und es fällt ihnen leicht, auf andere Menschen zuzugehen – was sie auch von sich aus tun.

Definition „*Verantwortungsbewusstsein*"

Verantwortungsbewusste Menschen übernehmen gerne Verantwortung für andere. Sie denken an die Konsequenzen ihres Handelns und die Folgen, die es für die Zukunft und für andere Menschen hat. Sie mischen sich (wenn nötig) auch in fremde Angelegenheiten ein, wenn ihre Einflussnahme zum Wohle anderer notwendig ist. Deshalb gehen sie sehr gewissenhaft vor, nehmen alles ernst und möchten sich selbst stets Rechenschaft ablegen können.

Jede Schülerin respektive jeder Schüler formulierte auf einer für jede Dimension erstellten Antworttabelle zu mindestens einer Dimension vier Verhaltensweisen. Diejenigen, welche vor Ablauf der Schulstunde mit der Beschreibung der Verhaltensweisen zu einer Dimension fertig waren, konnte noch zusätzlich eine zweite Dimension bearbeiten. Insgesamt formulierten die Schülerinnen und Schüler N = 248 Acts (Durchsetzungsfähigkeit n = 84, Kontaktfähigkeit n = 84, Verantwortungsbewusstsein n = 80).

2. Bearbeitung der generierten Acts

Aus den generierten Acts wählten die Studienleiterinnen diejenigen für die nachfolgende Einschätzung der Prototypizität aus, welche konkrete Verhaltensweisen in realen Situationen beschreiben und prinzipiell dem Erfahrungsschatz eines Jugendlichen entsprechen. Zudem wählten sie bei ähnlichen Verhaltensschilderungen jeweils nur eine aus. Die verbleibenden N = 149 Acts (Durchsetzungsfähigkeit n = 52, Kontaktfähigkeit n = 37 und Verantwortungsbewusstsein n = 60) brachten sie anschliessend in eine sprachlich einheitliche Form und stellten sie in einem Fragebogen zur Einschätzung der Prototypizität zusammen.

3. Einschätzung der mittleren Prototypizität

Zur Einschätzung der Prototypizität der generierten Acts bearbeiteten 19 Personen aus dem Bekanntenkreis der Studienleiterinnen und 20 Unteroffiziersschüler den Fragebogen mit den 149 nach den drei Dimensionen geordneten Acts anhand folgender Instruktion:

Auf den folgenden Seiten finden Sie 149 Aussagen, die konkrete Handlungen von Personen zu drei verschiedenen Eigenschaften beschreiben. Wir sind daran interessiert, zu erfahren, wie „typisch" Ihrer Meinung nach jede dieser Handlungen die Eigenschaft „Durchsetzungsfähigkeit", „Kon-

taktfähigkeit" oder „Verantwortungsbewusstsein" zum Ausdruck bringt. Zur begrifflichen Orientierung finden Sie zu jeder Eigenschaft eine Definition. Neben jeder Aussage befindet sich zur Einschätzung der Prototypizität eine 4-stufige Skala mit den Abstufungen „gar nicht typisch", „wenig typisch", „ziemlich typisch" und „sehr typisch". Schätzen Sie bitte für jede der geschilderten Handlungen die Typizität ein und kreuzen Sie das betreffende Feld an.

Um Personen aufzudecken, welche den Fragebogen unseriös ausgefüllt haben und ein vom Durchschnitt stark abweichendes Antwortverhalten zeigen – was vor allem bei den zur Befragung befohlenen Unteroffiziersschülern auf Grund mangelnder Motivation auftreten kann – führten wir eine Reliabilitätsanalyse über die einzelnen Rater durch. Dabei stellt die Korrelation r_{it} das Mass der Übereinstimmung der Angaben eines Raters mit dem Durchschnittsrating dar. Als Ausschlusskriterium definierten wir eine Grenze von r_{it} <.30, welche zwei Teilnehmer nicht erreichten (r_{it} =.08 resp. .24) womit für die Bestimmung der Prototypizität der einzelnen Acts noch N = 37 Einstufungen zur Verfügung standen.

Als Mass der Prototypizität der einzelnen Acts verwendeten wir den Mittelwert der Einstufungen. Die eingestuften Acts brachten wir pro Dimension in eine Rangfolge, um so die hochprototypischen unter ihnen als Ausgangsmaterial für die Formulierung der Item-Stämme zu verwenden. In den Tabellen 6.17 bis 6.19 sind pro Dimension beispielhaft jeweils fünf der hochprototypischen und drei der niedrigprototypischen Acts aufgeführt. Die vollständigen Tabellen sind im Anhang 6.11 bis 6.13 abgebildet. In der nachfolgend dargestellten Tabelle 6.16 sind die Verteilungskennwerte der Acts pro Dimension zusammenfassend dargestellt. Dabei wird ersichtlich, dass die mittlere Prototypizität der Acts der Dimension Durchsetzungsfähigkeit leicht tiefer ausfällt als diejenige der beiden anderen Dimensionen, wobei dieser Unterschied nicht signifikant ist ($\chi^2(2, N = 149) = 3.14$, $p = .21$)

Tabelle 6.16
Übersichtsdarstellung der Ergebnisse des Prototypenratings

Dimension	Anzahl Acts	mittlere Prototypizität	höchste Prototypizität	tiefste Prototypizität	Range der Standardabweichungen
		M	*M*	*M*	*SD*
Durchsetzungsfähigkeit	52	2.83	3.78	1.32	.53 bis 1.09
Kontaktfähigkeit	37	3.05	3.86	1.57	.35 bis 1.04
Verantwortungsbewusstsein	60	3.00	3.81	1.81	.48 bis 1.10

Anmerkung. N = 37. Die Skala reicht von 1 = „gar nicht typisch" zu 4 = „sehr typisch".

Tabelle 6.17

Hoch- und niedrigprototypische Acts zur Dimension Durchsetzungsfähigkeit

	Prototypizität	
	M	*SD*
Er setzte seine Ideen für die Schülerzeitung mit stichhaltigen Argumenten durch.	3.78	.53
Sie setzte ihren Vorschlag für eine neue Arbeitsaufteilung bei ihrem Chef durch.	3.68	.58
Er setzte seinen Wunsch, ins Kino zu gehen, gegen die Meinung seiner drei Kollegen durch.	3.62	.59
Sie verschaffte sich als Trainerin einer Männermannschaft Respekt, indem sie die Mannschaft durch gezieltes Training zum Erfolg führte.	3.50	.61
Sie sprach so lange auf ihren Lehrer ein, bis sie die misslungene Prüfung wiederholen durfte.	3.43	.83
Sie brachte mit ihren offenen Haaren und ihrem tiefen Kleidausschnitt die Männer dazu, ihr zuzuhören.	1.68	.82
Er war den ganzen Abend deprimiert, weil seine Freunde für einmal nicht das unternahmen, was er wollte.	1.59	.80
Er erniedrigte in einer Meinungsverschiedenheit seine Diskussionspartner und machte sich über deren Ansichten lustig.	1.32	.58

Anmerkung. *N* = 37. Die Skala reicht von 1 = „gar nicht typisch" zu 4 = „sehr typisch".

Tabelle 6.18

Hoch- und niedrigprototypische Acts zur Dimension Kontaktfähigkeit

	Prototypizität	
	M	*SD*
Sie nahm die Einladung von ihrem Freund, mit Bekannten von ihm ins Kino zu gehen, ohne zu zögern an und unterhielt sich sofort mit ihnen.	3.86	.35
Er setzte sich am ersten Schultag neben eine Mitschülerin und begann sofort ein Gespräch mit ihr.	3.81	.40
Er begann auf dem Markt in Italien ein Gespräch mit einer Frau und kümmerte sich nicht um die sprachlichen Hindernisse.	3.68	.47
Sie reiste ohne Holländischkenntnisse nach Amsterdam und hatte bereits nach zwei Tagen eine Arbeitsstelle und Freunde gefunden.	3.62	.59
Sie setzte sich in einem beinahe leeren Zug zu einer ihr unbekannten Person und begann ein Gespräch.	3.46	.61
Er bat im Standbad zwei unbekannte Frauen um etwas Sonnencreme.	2.30	.85
Sie entschloss sich spontan, etwas trinken zu gehen, weil die Schule am Nachmittag ausfiel.	1.76	.72
Er lachte, obwohl er den Witz nicht lustig fand.	1.57	.73

Anmerkung. *N* = 37. Die Skala reicht von 1 = „gar nicht typisch" zu 4 = „sehr typisch".

Tabelle 6.19

Hoch- und niedrigprototypische Acts zur Dimension Verantwortungsbewusstsein

	Prototypizität	
	M	*SD*
Er überprüfte die Schlucht, bevor er sie mit der Canyoning-Gruppe passierte.	3.81	.64
Er konsumierte an einer Party keinen Alkohol, da er mit dem Auto unterwegs war.	3.81	.57
Er kümmerte sich um den Verletzten, welcher regungslos am Waldrand lag.	3.59	.60
Sie half dem verzweifelten Kind in der Bahnhofshalle seine Mutter wiederzufinden.	3.57	.55
Er wies einen Teilnehmer der Wandergruppe, der den Weg verlassen wollte, auf die Gefahren seines Handelns hin.	3.41	.76
Er übernahm die Verantwortung für die Homepage des Badmintonclubs.	2.35	.82
Sie kaufte Bio-Fleisch, da sie kein Fleisch mit Antibiotika essen wollte.	2.24	.89
Sie übernahm die Verantwortung für ihre Kollegin, die bei der Prüfung von ihr abschrieb.	1.81	.84

Anmerkung. *N* = 37. Die Skala reicht von 1 = „gar nicht typisch" zu 4 = „sehr typisch".

Diese Listen mit den nach der eingeschätzten Prototypizität rangierten Acts lieferten uns zuerst einmal wertvolle Hinweise für die Anpassung der Definitionen der drei im Leadership-Fragebogen enthaltenen Verhaltensdimensionen.

Für die Acts, welche wir als Bestandteile der Item-Stämme einsetzten, wählten wir aus den hochprototypischen diejenigen aus, welche sich als Ausgangsmaterial für anhand des Wertequadrates entwickelte Items am besten eignen. Wir formulierten die so ausgewählten Acts um und kleideten sie in einen gut definierten Situationskontext ein. Dabei strebten wir das Ziel an, dass sich jeder Testbearbeiter problemlos in die geschilderten Situationen hineindenken und –fühlen kann. Somit haben wir – im Gegensatz zum klassischen Vorgehen beim AFA – den Situationskontext deutlich hervorgehoben, indem wir spezifische Situationen aus dem Leben Jugendlicher als „Stimulus" verwendeten. Nachfolgend verdeutliche ich diese Vorgehensweise anhand von zwei Beispielen:

Das erste Beispiel zeigt auf, wie wir einen Act praktisch wortwörtlich in ein entsprechendes Leadership-Fragebogen-Item übernommen haben. Dazu schildern wir im Item-Stamm eine entsprechende Situation und übernehmen den eigentlichen Act, also die konkret gezeigte Verhaltensweise, als eine der vier möglichen Antwortalternativen des Items (Moroge & Schibli, 2002). (Das Vorgehen bei der Entwicklung der Antwortalternativen stelle ich im nächsten Kapitel dar.)

Act:	Er nahm seinem betrunkenen Freund die Autoschlüssel weg.
Item-Stamm:	Sie sind mit ein paar Kollegen in einer Disco. Einer von ihnen verabschiedet sich und möchte mit seinem Auto nach Hause fahren. Sie haben allerdings den Eindruck, dass er zuviel getrunken hat, um noch sicher Auto fahren zu können.
Antwortalternative:	Sie nehmen Ihrem Kollegen die Autoschlüssel weg, nötigenfalls auch gegen seinen Willen.

Das zweite Beispiel zeigt auf, wie wir Elemente des Acts angepasst haben, um den Aufforderungscharakter der Situation deutlicher hervorzuheben.

Act:	Sie riet ihrer Freundin, die über Schmerzen klagte, zum Arzt zu gehen.
Item-Stamm:	Ihnen fällt auf, dass eine Kollegin von Ihnen immer dünner wird. Sie haben gehört, dass sie seit längerem nach den Mahlzeiten erbricht, um ihr Gewicht zu halten.
Antwortalternative:	Sie raten Ihrer Kollegin, einen Arzt oder eine Beratungsstelle aufzusuchen.

Weiter entwickelten wir auch Items, bei welchen wir mehrere Acts zusammenfassten oder uns von verschiedenen Acts zur Formulierung einer neuen Situation inspirieren liessen. Die Umformulierung und Umgestaltung der mit dem AFA gesammelten Acts hat zur Folge, dass die Prototypenratings für die Leadership-Fragebogen-Items keine Gültigkeit mehr haben. Auf Grund des gewählten Antwortformates lässt sich einem Item auch nicht mehr ein Prototypen-Rating-Wert zuteilen, sondern dieser müsste für jede Antwortalternative einzeln erfolgen. Wenn man die von Angleitner, Buss und Demtröder (1990) geforderte Mehrfachsortierung (*multiple dispositional act sorting*) einsetzt, würde das bedeuten, dass Probanden für jedes von den in der ersten Version des Leadership-Fragebogens enthaltenen 39 Items für jede der vier Antwortkategorien und jede der drei Dimensionen eine Prototypeneinschätzung vornehmen müssten, insgesamt also 468. Das gewählte Antwortformat führt auch dazu, dass beim Leadership-Fragebogen eine andere Auswertung vorzunehmen ist, als dies Buss und Craik (1980) vorschlagen: Bei einer AFA-Skala hat der Bearbeiter anzugeben, wie oft er ein bestimmtes Verhalten zeigt (Aufsummierung: Wie häufig verhält er sich dominant?), beim Leadership-Fragebogen – wie bei SJT üblich – wie stark ausgeprägt er dieses Verhalten ausführt (Gewichtung: Wie dominant verhält er sich?).

Um den Fragebogen für die Stellungspflichtigen möglichst attraktiv zu gestalten, suchte ich in Fotosharing-Communities im Internet zu jeder Situation ein passendes Bild. Diese Art der Itempräsentation übernahm ich von Etzel (1999), der sie in seinem profacts-Testsystem einsetzt (siehe Kap. 1.3, Abb. 1.2).

Zur Erstellung des Itempools für die erste Datenerhebung im Rahmen der Testentwicklung mussten wir in einem zweiten Schritt mögliche Verhaltensalternativen zu den in den Item-Stämmen geschilderten Situationen konstruieren. Dies geschieht in Anlehnung an die Konstruktionsweise von SJTs, da bei einer rein nach dem AFA entwickelten Skala die Situation und die Verhaltensweise eine Einheit bilden, womit sich die zusätzliche Entwicklung von Verhaltensalternativen erübrigt. Bei einem SJT übernehmen diese Aufgabe die Testentwickler oder diese befragen Subject Matter Experts (Weekley, Ployhart & Holtz, 2006). Wir wählten als Konstruktionsrational für die Formulierung der Verhaltensalternativen – wie ich im nächsten Kapitel erläutere – das Wertequadrat.

6.3 Entwicklung der Wertequadrate der Testdimensionen

Als Konstruktionsrational (siehe Kapitel 1.3) für die Generierung der Verhaltensrespektive Antwortalternativen der Leadership-Fragebogen-Items wählten wir das Wertequadrat von Helwig (1948; siehe auch Schulz von Thun, 1989; Westermann, 2007a). Damit wollte ich ein Regelwerk erstellen, anhand dessen wir die jeweiligen Antwortalternativen formulieren können, damit diese über alle Items einer Dimension vergleichbar sind.

Von den in Kapitel 4.3 dargestellten Vorgehensweisen für die Entwicklung von Wertequadraten hat sich das von Gloor (2007) als am ehesten zielführend herausgestellt. Bei den von Westermann (2007b) vorgestellten fünf Denkschritten führt deren Transformation in einen Entwicklungsprozess zu einer Überdeterminierung und somit zu einem unnötig grossen Aufwand. Schulz von Thun (1989) geht bei seiner Entwicklungsanweisung der Gedankenlogik des Wertequadrates nach, indem er zur Übertreibung die gegenteilige Tugend, welche zugleich den positiven Gegenwert zur Ausgangstugend darstellt, sucht. Würde, wie Gloor es vorschlägt, zuerst noch die zweite Übertreibung definiert, liesse sich die noch fehlende zweite Tugend jedoch genauer definieren, da man sie quasi als fehlendes Puzzlestück jetzt genau in das Wertequadrat einpassen muss.

In Abbildung 6.2 habe ich Gloors Vorgehensweise mit der für unsere Rahmenbedingungen angepassten Terminologie schematisch dargestellt: So unter-

scheide ich die beiden Tugenden in eine erwünschte und eine alternative Verhaltensweise. Mit dieser Bevorzugung werden wir einer der Grundannahmen des Wertequadrates – derjenigen der Gleichwertigkeit der beiden Tugenden – nicht ganz gerecht. Da mit unseren Wertequadraten jedoch auch ein Anforderungsprofil verknüpft ist, welches die erfolgsrelevanten Verhaltensweisen aufzeigt, drängt sich eine solche Gewichtung automatisch auf.

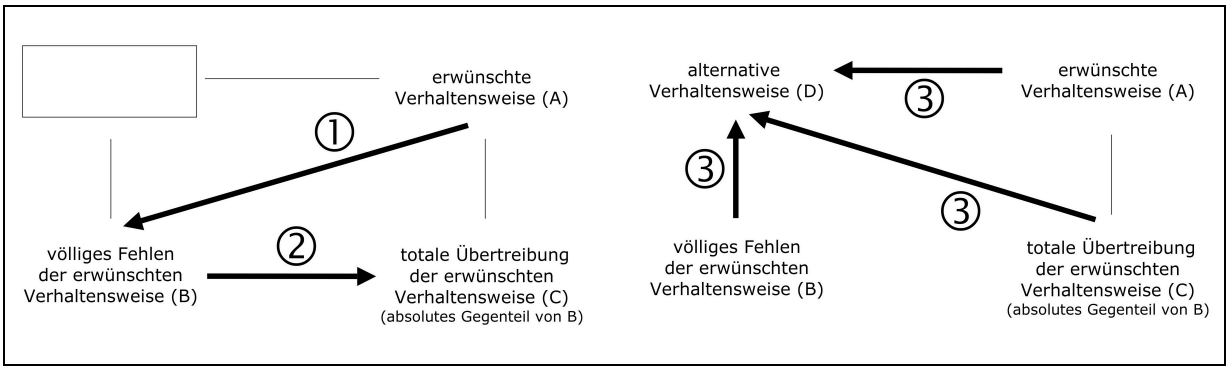


Abbildung 6.2 Vorgehen beim Definieren der vier Wertequadranten.

Bei der Erstellung der Wertequadrate setzen wir jeweils den Namen der Leadership-Dimension an die Stelle der erwünschten Verhaltensweise. Wir starten also den Erstellungsprozess, indem wir in einem Wertequadrat-Grundgerüst bei der Tugend oben rechts die Dimensionsbezeichnung – zum Beispiel den Begriff Durchsetzungsfähigkeit – notieren. Davon ausgehend definieren wir in einem ersten Schritt den Begriff, welcher das völlige Fehlen der gewünschten Eigenschaft beschreibt – im gewählten Beispiel die Selbstverleugnung – und setzen ihn in das diagonal gegenüberliegende Feld der erwünschten Verhaltensweise, an die Stelle der einen Übertreibung. Im zweiten Schritt bestimmen wir die Bezeichnung für die Übertreibung der erwünschten Verhaltensweise, also das totale Gegenteil des eben definierten Begriffes. Beim Wertequadrat zur Dimension Durchsetzungsfähigkeit haben wir uns dabei für Rücksichtslosigkeit entschieden. Anhand aller drei bisher gewählten Bezeichnungen suchen wir abschliessend nach der noch fehlenden Tugend, der alternativen Verhaltensweise, welche eine positive Eigenschaft beschreibt, als Referenzpunkt für die eine Übertreibung passt und im Gegensatz zur zweiten Übertreibung steht. Bezüglich Durchsetzungsfähigkeit entschieden wir uns für die Kompromissfähigkeit.

Auf den nächsten Seiten stelle ich die auf diese Weise erstellten drei Wertequadrate des Leadership-Fragebogens mit den entsprechenden Definitionen der vier Wertequadranten vor (Moroge & Schibli, 2002).

Wertequadrat Durchsetzungsfähigkeit

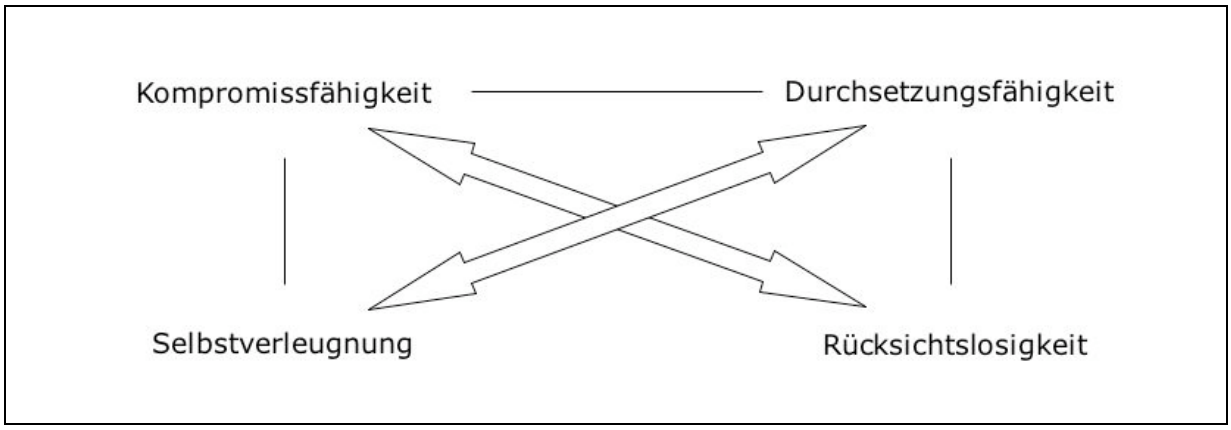


Abbildung 6.3 Das Wertequadrat zur Dimension Durchsetzungsfähigkeit.

Selbstverleugnende Menschen vernachlässigen ihre eigenen Bedürfnisse und Interessen. Auf Grund eines starken Harmoniebedürfnisses vermeiden sie jede Konfrontation und gehen Konflikten aus dem Weg. Sie vertreten keinen eigenen Standpunkt oder geben diesen, wie auch ihre Ziele, bei Widerständen vorschnell auf. Sie ordnen sich ihrem sozialen Umfeld völlig unter.

Kompromissfähige Menschen berücksichtigen die Interessen anderer ebenso wie die eigenen. Sie beharren nicht um jeden Preis auf ihrem Standpunkt, sondern sind bemüht, partnerschaftliche Lösungen zu suchen. Bei Widerständen versuchen sie, ihren Standpunkt wie auch ihre Ziele mit denjenigen anderer abzugleichen und sind zu Zugeständnissen bereit. Dabei sind sie diplomatisch und zeigen Verhandlungsgeschick.

Durchsetzungsfähige Personen berücksichtigen die eigenen Interessen stärker als diejenigen anderer. Sie versuchen ihren Standpunkt gegenüber anderen zu wahren, indem sie diesen mit Nachdruck vertreten und vermitteln. Sie verfolgen auch dann hartnäckig ihre Ziele, wenn sich Widerstände ergeben und überwinden diese, indem sie mit Argumenten überzeugen. In Diskussionen sind sie offensiv, selbstbewusst und scheuen sich nicht, sich dabei zu exponieren.

Rücksichtslose Menschen berücksichtigen nur ihre eigenen Interessen und wirken dadurch egoistisch. Sie sind nicht bereit, Kompromisse einzugehen oder ihren Standpunkt aufzugeben auch wenn sie dafür von ihrem sozialen Umfeld negative Beurteilungen in Kauf nehmen müssen. Sie setzen sich über Widerstände kompromisslos hinweg und verwirklichen so ihre Ziele ohne sich um Verluste zu kümmern oder Rücksicht auf andere zu nehmen.

Wertequadrat Kontaktfähigkeit

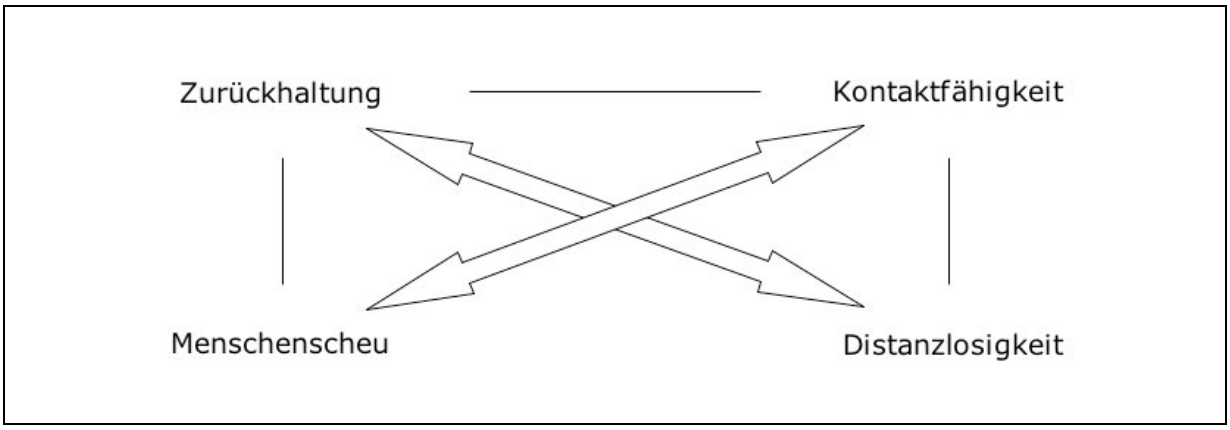


Abbildung 6.4 Das Wertequadrat zur Dimension Kontaktfähigkeit.

Menschenscheue Personen sind am liebsten allein und ungern in Gesellschaft. Sie vermeiden es, Kontakt zu anderen aufzunehmen und sind unzugänglich. Sie wirken verschlossen, sind Einzelgänger und haben kaum Freunde und wirken im Umgang mit anderen gehemmt.

Zurückhaltende Menschen sind lieber allein als in Gesellschaft. Bei der Kontaktaufnahme verhalten sie sich passiv und überlassen ihrem Gegenüber die Initiative. Beim Eingehen neuer Beziehungen sind sie vorsichtig und selektiv. Sie sind im Umgang mit anderen eher abwartend und zurückhaltend und wirken dadurch reserviert.

Kontaktfähige Menschen sind lieber in Gesellschaft als allein. Die Kontaktaufnahme mit fremden Menschen fällt ihnen leicht und im Umgang mit Mitmenschen sind sie unbefangen und offen. Sie sind sehr darum bemüht, Freundschaften zu schliessen und aufrechtzuerhalten und verfügen über ein grosses Beziehungsnetz.

Distanzlose Menschen fühlen sich nur in Gesellschaft wohl und sind nicht in der Lage, allein zu sein. Sie haben ein ausgeprägtes Mitteilungsbedürfnis und ein zwanghaftes Bedürfnis nach Anschluss. Oft haben sie eine Unmenge an Bekannten aber keine wirklichen Freundschaften. In der Kontaktaufnahme verhalten sie sich aufdringlich und wahllos. Sie sind im Umgang mit anderen oberflächlich, indiskret und egozentrisch und wirken zuweilen respektlos.

Wertequadrat Verantwortungsbewusstsein

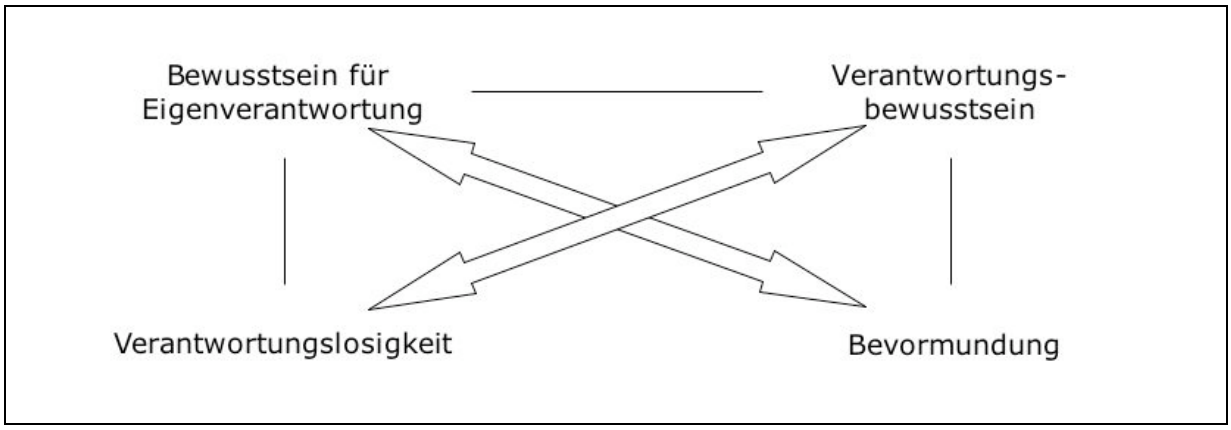


Abbildung 6.5 Das Wertequadrat zur Dimension Verantwortungsbewusstsein.

Verantwortungslose Menschen übernehmen keine Verantwortung für andere. Sie kümmern sich prinzipiell nicht um fremde Angelegenheiten und sind der Auffassung, dass jeder für sich selbst verantwortlich ist. Dies kommt in einer passiven, gleichgültigen Haltung dem sozialen Umfeld gegenüber zum Ausdruck.

Menschen mit *Bewusstsein für Eigenverantwortung* übernehmen nur in einem eng gefassten Bereich Verantwortung. Sie kümmern sich nur um Dinge, die sie unmittelbar etwas angehen und mischen sich nicht in fremde Angelegenheiten ein. Sie fördern einerseits die Eigenverantwortung ihres sozialen Umfeldes, andererseits erkennen sie die Grenzen ihres Zuständigkeitsbereiches an.

Verantwortungsbewusste Menschen nehmen Verantwortung wahr und fühlen sich verpflichtet, für andere Verantwortung zu übernehmen. Sie kümmern sich wenn nötig auch um fremde Angelegenheiten. Dies kommt in einer fürsorglichen Haltung dem sozialen Umfeld gegenüber zum Ausdruck.

Bevormundende Menschen übernehmen die Verantwortung für andere ohne deren Autonomie zu respektieren. Sie erteilen ungefragt Ratschläge, schreiben anderen vor, was diese tun oder lassen sollen und beeinflussen diese in ihren Entscheidungen und mischen sich so ungefragt und unnötig in fremde Angelegenheiten ein. Dies kommt in einer anmassenden Haltung dem sozialen Umfeld gegenüber zum Ausdruck.

6.4 Entwicklung der Items und der Testendform des Leadership-Fragebogens

Mit den anhand des Act Frequency Approachs generierten Situationsschilderungen bildeten zwei studentische Arbeitsgruppen (Bühler Ruedin & Selk, 2001; Moroge & Schibli, 2002) Item-Stämme, zu welchen wir anschliessend mit Hilfe der Wertequadrate die Verhaltensalternativen entwickelten. Nachfolgend ein Beispiel zur Dimension Durchsetzungsfähigkeit.

38. Schülerzeitung

Sie arbeiten für die Schülerzeitung und haben einen Artikel über ein wichtiges und aktuelles Thema verfasst. Weil es nur noch für einen Artikel Platz hat, kommt es zu einer Auseinandersetzung mit einem Kollegen, der seinen Artikel wichtiger findet.

- ☐ Sie haben keine Lust auf stundenlange Diskussionen. Dann wird Ihr Beitrag eben in einer der nächsten Ausgaben erscheinen.
- ☐ Da keiner nachgeben will, schlagen Sie vor, dass beide ihren Beitrag so kürzen, dass zwei Artikel Platz haben.
- ☐ Sie lassen nicht locker und versuchen, die anderen Redaktionsmitglieder von der Wichtigkeit Ihres Beitrages zu überzeugen.
- ☐ Sie lassen der Redaktion keine Wahl: Entweder Ihr Artikel wird abgedruckt oder Sie werden in Zukunft nicht mehr für diese Zeitung arbeiten.

Abbildung 6.6 Item aus einer Vorversion des Leadership-Fragebogens zur Dimension Durchsetzungsfähigkeit.

Bei der Testentwicklung gingen wir iterativ vor, indem wir die Items in einem Pretests überprüften und anschliessend auf der Grundlage der erhaltenen Ergebnisse die Items überarbeiteten und neue Items entwickelten. Insgesamt legten wir im Rahmen von Pretests vier unterschiedliche Vorversionen des Leadership-Fragebogens mit insgesamt 81 Items 643 Rekruten aus fünf verschiedenen Schulen vor. Dabei war es von grossem Vorteil, dass die Rekruten praktisch identisch mit der späteren Zielpopulation – den diensttauglichen Stellungspflichtigen – sind. Ein deutlicher Unterschied lässt sich jedoch bei der Motivation, den Test zu bearbeiten, ausmachen: Obwohl es in beiden Populationen vorkommt, dass Einzelne den Fragebogen ungenau und unseriös ausfüllen, tritt die bewusste Eindrucksteuerung – wie in Kapitel 1.1 dargestellt – nur in der Selektionssituation bei den Stellungspflichtigen auf. An den nachfolgenden Überprüfungsstudien mit der ersten definitiven Version des Fragebogens mit 39 Aufgaben nahmen weitere

210 Rekruten und 124 Unteroffiziere aus insgesamt acht verschiedenen Rekrutenschulen teil. In Tabelle 6.20 sind die einzelnen Studien im Überblick aufgeführt.

Tabelle 6.20
Stichproben der Pretestung und ersten Überprüfung des Leadership-Fragebogens

Datenerhebung[2]	Rekrutenschule	Stichproben-umfang	Anzahl Items
Bühler Ruedin & Selk (2001)	Übermittlungspioniere	119	54
Moroge & Schibli (2002)	Flieger-Bodentruppen	85	46
	Panzergrenadiere	153	60
	Infanterie, Artillerie	286	48
Bauhofer, Bösser & Henggeler (2003)	Flieger-Bodentruppen, Übermittlungs-pioniere, Infanterie-Durchdiener	210	39
Imper & Maier (2003)	Unteroffiziere der Artillerie, Flieger-abwehr, Panzer, Infanterie, Genie	124	39

Die erste definitive Version des Leadership-Fragebogens bestand aus 13 Items pro Dimension und wurde 2003 mit einer einfach handhabbaren Einzelplatzsoftware (*Teach and Test*, TNT) auf den Computern in den Rekrutierungszentren der Schweizer Armee eingesetzt. In der zwei Jahre später auf dem für den Bedarf in den Rekrutierungszentren der Schweizer Armee angepassten Testadministrationssystem *Computer-Assistiertes Testen* (CAT) implementierten Version des Fragebogens fügte ich – wie in Abbildung 6.7 dargestellt – die die einzelnen Situationen illustrierende Fotografien hinzu.

Abbildung 6.7 Itemlayout der beiden Computerversionen TNT und CAT.

[2] Eine Übersicht über alle verwendeten Stichproben ist am Schluss der Arbeit aufgeführt.

Anhand der 2003 innerhalb eines knappen Jahres in den Rekrutierungszentren erhobenen Daten verkürzte ich den Fragebogen auf zehn Items pro Skala. Insgesamt standen mir dafür die Datensätze von 10'631 deutschschweizer Stellungspflichtigen zur Verfügung. Von den 7'887 vollständigen Datensätzen zum Leadership-Fragebogen schloss ich diejenigen Stellungspflichtigen aus (n = 16), welche in einem Item eine Bearbeitungszeit von unter 10 Sekunden aufwiesen, was auf eine unseriöse Bearbeitung des Fragebogens schliessen lässt. Somit umfasste die endgültige Stichprobe 7'871 Stellungspflichtige.

Für die weiteren Analysen habe ich ein Scoring verwendet, welches bei der jeweiligen Antwortalternative den Grad der Ausprägung in der zu messenden Eigenschaft widerspiegelt: Der niedrigsten Ausprägung weise ich den Wert 1 zu – für die Dimension Durchsetzungsfähigkeit entspricht dies der Antwortalternative Selbstverleugnung –, der höchsten den Wert 4 – im gewählten Beispiel also der Antwortalternative Rücksichtslosigkeit. Die drei Skalenwerte bilde ich durch reine Aufsummierung der Scorings der der jeweiligen Skala zugeordneten Items.

In den Tabellen 6.22 bis 6.24 sind die Itemkennwerte der ursprünglichen und der gekürzten Skalen aufgeführt. Der Vergleich der Cronbach Alphas der einzelnen Skalen zeigt auf, dass die Kürzung um immerhin knapp 25% kaum einen Einfluss auf die Homogenität respektive Reliabilität der Skalen hat. Weiter zeigen sich auch die grossen Unterschiede in der Homogenität zwischen den einzelnen Skalen: Einzig die Skala Kontaktfähigkeit lässt sich mit einem α = .83 als zufrieden stellend bezeichnen. Eine ungenügende Reliabilität zeigt sich hingegen bei der Skala Durchsetzungsfähigkeit mit einem α = .53. Die Skala Verantwortungsbewusstsein liegt mit einem α = .71 gemäss der DIN *„Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen*" (2002) noch knapp im tolerierbaren Bereich. Auch Rückert (1993, S. 37) gibt „0.70 als unteren Grenzwert für hinreichend zuverlässige individualdiagnostische Aussagen" an. Nunnally und Bernstein (1994) sehen beim Einsatz eines Tests als Grundlage für wichtige Entscheidungen jedoch eine Reliabilität von .90 als absolutes Minimum und eine solche von .95 „should be considered the desirable standard" (S. 265). Diese Forderung relativieren Moosbrugger und Rauch (2005, S. 183) mit ihrer Aussage, dass „Werte grösser als 0.85 wünschenswert, aber nur schwer realisierbar und deshalb selten" sind. In der nachfolgenden Tabelle führe ich die Bewertungsrichtlinien der *European Federation of Psychologists' Associations* (EFPA; Lindley, Bartram & Kennedy, 2008) und des *Committee on Test Affairs Netherlands* (COTAN; Evers, 2001) auf. Das COTAN unterscheidet dabei – basierend auf Nunnally und Bernstein – zwischen Testeinsätzen, von welchen weniger wichtige oder wichtige Entscheidungen abhängen.

Tabelle 6.21

Richtlinien für die Bewertung der Reliabilität (internale Konsistenz)

Reliabilität	EFPA	COTAN (weniger wichtige Entscheidungen)	COTAN (wichtige Entscheidungen)
$r < .70$	unangemessen	ungenügend	–
$.70 < r < .80$	angemessen	genügend	ungenügend
$.80 < r < .90$	gut	gut	genügend
$r > .90$	exzellent	–	gut

Wenn wir in Betracht ziehen, dass die Psychologen der Rekrutierungszentren den Leadership-Fragebogen als eines von mehreren Instrumenten im Rahmen einer Selektion mit Select-out-Charakter einsetzen und die schlussendlich vom Rekrutierungsoffizier abgegebene Kaderempfehlung für die Auftraggeber – also die Kommandanten der Rekrutenschulen – nicht zwingend verbindlich ist, so ist der Einfluss des Leadership-Fragebogens auf eine zukünftige Kaderlaufbahn eher gering. Diese Tatsachen rechtfertigen meines Erachtens die Berücksichtigung des Reliabilitätskriteriums von $r = .70$.

Tabelle 6.22

Vergleich der Itemkennwerte der ursprünglichen und der gekürzten Version der Skala Durchsetzungsfähigkeit

Item	13-Item-Version				10-Item-Version	
	M	*SD*	r_{it}	Faktor-ladung	r_{it}	Faktor-ladung
Disco	2.03	.99	.20	.38		
Lohnerhöhung	2.42	.86	.23	.40	.20	.39
Fahrer	2.44	.75	.25	.45	.25	.48
Unterbruch	2.00	.79	.26	.44	.24	.44
Zugreise	2.45	.74	.14	.26		
Aufräumen	2.61	.72	.27	.46	.26	.47
Waschküche	1.88	.60	.26	.45	.25	.47
Geschirr	2.72	.85	.23	.40	.23	.43
Musik	2.32	.78	.20	.38	.19	.39
Probleme	1.90	.97	.21	.38	.21	.41
Auswärts	2.53	.89	.21	.39	.23	.44
Arbeit	2.61	.87	.22	.40	.23	.44
Schülerzeitung	1.99	.80	.19	.33		
Cronbach Alpha der Skala			.56		.53	

Anmerkung. N = 7'871. r_{it} = Trennschärfe.

Tabelle 6.23

Vergleich der Itemkennwerte der ursprünglichen und der gekürzten Version der Skala Kontaktfähigkeit

Item	13-Item-Version				10-Item-Version	
	M	*SD*	r_{it}	Faktor-ladung	r_{it}	Faktor-ladung
Zelten	2.71	.82	.36	.46		
Nachbarn	2.89	.74	.53	.53	.54	.55
Zugfahrt	2.69	.71	.61	.69	.60	.70
Kurs	2.56	.77	.56	.56	.54	.58
Schultag	2.77	.93	.49	.54	.49	.57
Flugzeug	2.77	.88	.57	.62	.57	.65
Barmann	2.27	.85	.45	.58	.45	.60
Begleitung	2.37	.89	.52	.61	.51	.63
Party	2.17	.75	.40	.50		
Alleingelassen	2.56	.69	.44	.52		
Umzug	2.92	.77	.48	.49	.48	.49
Geburtstag	2.45	.76	.52	.60	.50	.60
Fitness	2.61	.87	.53	.60	.53	.63
Cronbach Alpha der Skala			.84		.83	

Anmerkung. N = 7'871. r_{it} = Trennschärfe.

Tabelle 6.24

Vergleich der Itemkennwerte der ursprünglichen und der gekürzten Version der Skala Verantwortungsbewusstsein

Item	13-Item-Version				10-Item-Version	
	M	*SD*	r_{it}	Faktor-ladung	r_{it}	Faktor-ladung
Schanze	2.63	.86	.30	.47	.31	.50
Unstimmigkeiten	2.47	.87	.35	.38		
Beratungsstelle	2.59	.67	.35	.45	.34	.46
Silvester	3.03	.90	.41	.50	.40	.51
Subvention	2.31	.77	.37	.40	.35	.40
Nachhilfestunden	2.45	.72	.41	.49	.40	.51
Kind	2.75	.78	.37	.45	.36	.46
Malediven	2.51	.76	.23	.34		
Wohnung	2.73	.78	.35	.48	.33	.48
Bergtour	2.98	.88	.40	.51	.38	.52
Ampel	2.91	.96	.43	.54	.43	.54
Autofahren	3.05	.85	.39	.50	.39	.53
Mädchen	2.68	.59	.32	.34		
Cronbach Alpha der Skala			.74		.71	

Anmerkung. N = 7'871. r_{it} = Trennschärfe.

In Tabelle 6.25 habe ich die Korrelationen zwischen den drei Skalen der 13- und der 10-Item-Version des Leadership-Fragebogens dargestellt. Sie fallen für beide Testversionen in etwa gleich aus und zeigen auf, dass die Skala Durchsetzungsfähigkeit keinen Zusammenhang mit den beiden anderen Skalen hat. Hingegen ist die Korrelation zwischen den Skalen Kontaktfähigkeit und Verantwortungsbewusstsein mit r = .56 respektive r = .53 so hoch, dass nicht mehr von unabhängigen Konstrukten gesprochen werden kann. Der Vergleich der mit dem NEO-Persönlichkeitsinventar (Ostendorf & Angleitner, 2004) operationalisierten Big Five zeigt jedoch, dass auch zwischen von der Theorie her unkorrelierten Faktoren mittlere Korrelationen auftreten können: So korreliert Offenheit mit Extraversion mit r = .40 und Neurotizismus mit Gewissenhaftigkeit mit r = -.37 (siehe Tabelle 6.26). Auch zwischen den 14 Skalen des Bochumer Inventars zur berufsbezogenen Persönlichkeitsbeschreibung (BIP, Hossiep & Paschen, 2003) treten zum Teil hohe Interkorrelationen auf. Die Spannbreite liegt zwischen r = -.34 und r = .75 und insgesamt erreichen 17 der 91 Interkorrelationen – also knapp 20% – einen Wert über r = .50. Beispielhaft führe ich hier die Korrelationen von Kontaktfähigkeit mit Selbstbewusstsein – r = .55 – und mit Sensitivität – r = .54 – auf.

Tabelle 6.25

Korrelationen zwischen den drei Skalen des Leadership-Fragebogens

	13-Item-Version		10-Item-Version	
	Kontakt-fähigkeit	Verant-wortungs-bewusstsein	Kontakt-fähigkeit	Verant-wortungs-bewusstsein
Durchsetzungsfähigkeit (13/10)	.04***	.05***	-.00	.04***
Kontaktfähigkeit (13/10)		.56***		.53***

Anmerkung. N = 7'871.

Tabelle 6.26

Korrelationen zwischen den Skalen des NEO-PI-R (nach Ostendorf & Angleitner, 2004, S. 104 und 109)

	N	E	O	A	C
Neurotizismus	*.92 (.71)*				
Extraversion	-.27	*.89 (.63)*			
Offenheit	.05	.40	*.89 (.63)*		
Verträglichkeit	-.04	-.05	.02	*.87 (.58)*	
Gewissenhaftigkeit	-.37	.08	-.10	.06	*.90 (.65)*

Anmerkung. *In der Diagonale sind die Reliabilitäten (Cronbach Alpha) und in Klammern die nach der Spearman-Brown-Formel auf zehn Items korrigierten Werte aufgeführt.*

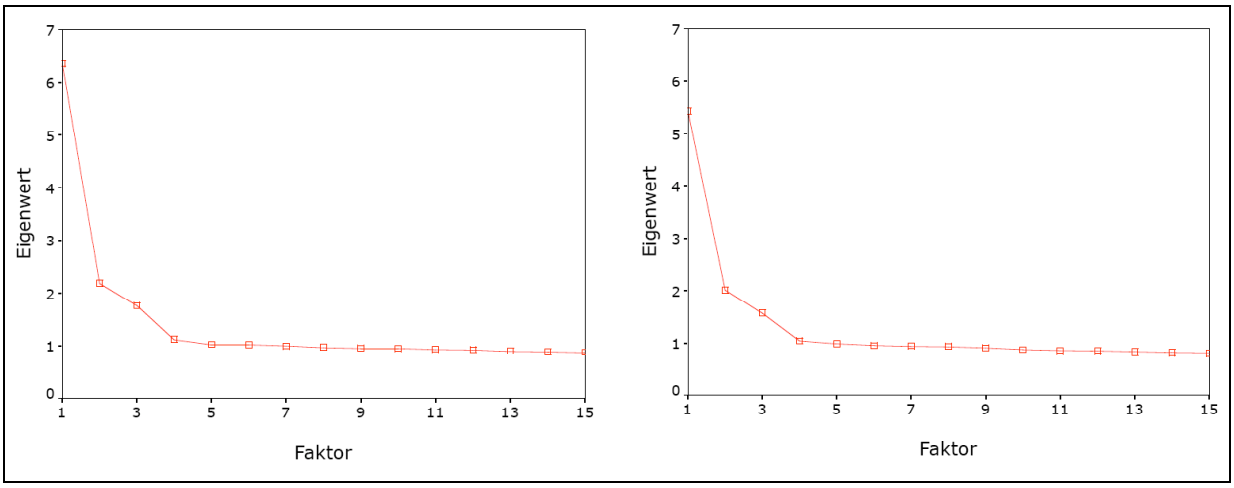


Abbildung 6.8 Scree-Plots der Faktorenanalysen der 39- und der 30-Item-Versionen des Leadership-Fragebogens.

Zur Bestimmung der Faktorenstruktur führte ich mit den 39 respektive 30 Items des Leadership-Fragebogens Hauptkomponentenanalysen mit orthogonaler Rotation (Varimax) durch, welche jeweils in fünf Iterationen konvergierten. Ich entschied mich trotz der Korrelation der beiden Skalen Kontaktfähigkeit und Verantwortungsbewusstsein für eine orthogonale Rotation, da ich bei der Testkonstruktion das Ziel verfolgte, drei voneinander unabhängige Skalen zu entwickeln. Die Voraussetzungen für die Durchführung einer Faktorenanalyse sind gegeben: Das Mass der Stichprobeneignung nach Kaiser-Meyer-Olkin beträgt *KMO* = .94 respektive *KMO* = .93, was nach Kaiser (1974) als ‚fabelhaft' zu beurteilen ist. Zudem sind alle KMO-Werte der einzelnen Variablen grösser als .61 respektive .62 und liegen somit über der Grenze von .50. Der Bartlett-Test auf Sphärizität wird signifikant ($\chi^2(741, N = 7'871) = 43'697.70$, $p < .001$ resp. $\chi^2(435, N = 7'871) = 34'223.86$, $p < .001$), was bedeutet, dass die Korrelationen zwischen den Items genügend hoch für eine Hauptkomponentenanalyse sind. Die Determinanten der Korrelationsmatrizen betragen $|R| = 0.00384$ respektive $|R| = 0.01285$ und liegen somit über dem Grenzwert von .00001 (Field, 2009). Die Tests nach Haitovsky (1969, siehe auch Rockwell, 1975) werden nicht signifikant ($\chi^2_H(741, N = 7'871) = 30.23$, $p > .05$ resp. $\chi^2_H(435, N = 7'871) = 101.64$, $p > .05$), womit Multikollinearität vorliegen könnte. Zur Bestimmung der Anzahl Faktoren verwendete ich den Scree-Plot, welcher bei mehr als 200 Probanden als gutes Kriterium gilt (Stevens, 2002). Er zeigt – wie in Abbildung 6.8 ersichtlich – klar drei Faktoren auf, welche zusammen 26.46% resp. 30.01% der Varianz erklären, was deutlich unter der von Kline (1998) geforderten Grenze von 70% liegt. Diese ist jedoch sehr hoch angesetzt: Andere Autoren schlagen Werte zwischen 50% und 60% und mindestens 5% pro Faktor vor (z. B. Netemeyer, Bearden & Shar-

ma, 2003). Diese Anforderungen scheinen auch realistisch zu sein: So liegt die erklärte Varianz beim Fragebogen zu Kompetenz- und Kontrollüberzeugungen bei 41% (Krampen, 1991), beim NEO-PI-R bei knapp 60% (Ostendorf & Angleitner, 2004) oder beim Leistungsmotivationsinventar bei 63% (Schuler, Prochaska & Frintrup, 2001).

In den Anhängen 6.14 und 6.15 sind die Faktorladungen nach der Rotation aufgeführt. Alle Items laden auf den entsprechenden Faktor ohne bedeutsame Nebenladungen auf die jeweils beiden anderen Faktoren mit Ausnahme der Kontaktfähigkeitsitems „Nachbarn" und „Umzug" welche Nebenladungen zwischen .30 und .35 auf den Faktor Verantwortungsbewusstsein haben.

Mit den Daten aus den Rekrutierungszentren des Jahres 2003 führte ich auch die erste umfassende Normierung des Leadership-Fragebogens durch. In die Normierungsstichprobe für die deutschsprachige Version des Leadership-Fragebogens nahm ich nur diejenigen Stellungspflichtigen auf, zu welchen ich über einen vollständigen Datensatz verfügte, insgesamt 9'906. Dies stellt nach Lindley et al. (2008) eine sehr grosse Stichprobe dar und garantiert somit sehr verlässliche Normen. Die Normierung führte ich anhand einer nicht-linearen Transformation über den Prozentrang (Flächentransformation) in die Stanine-Norm (μ = 5; σ = 2) durch (Lienert & Raatz, 1998). Der Einheitlichkeit folgend, benütze ich für die hier aufgeführte Normierung den bisher verwendeten, 7'871 Stellungspflichtige umfassenden Datensatz, was zu geringfügigen Abweichungen in den Extrembereichen zu den in den Rekrutierungszentren eingesetzten Normen führt. In Tabelle 6.27 sind die Verteilungskennwerte, in Abbildung 6.9 die Verteilungen der Rohwert-Scores und der Stanine-Werte der drei Skalen des Leadership-Fragebogens dargestellt. Die Prozentrangverteilungen mit den zugeordneten Stanine-Werten sind im Anhang 6.16 aufgeführt. Ursachen für die augenfälligen Unterschiede in den Rohwert-Verteilungen können einerseits ungleich extrem definierte Übertreibungen der jeweiligen Wertequadrate oder aber das Bild, das sich die Stellungspflichtigen von einem militärischen Vorgesetzten machen, sein. Letztere Annahme weiterverfolgend, liesse sich eine militärische Führungsperson wie folgt beschreiben: Sie weist eine mittlere Ausprägung in der Durchsetzungsfähigkeit auf, ohne in die Extreme abzugleiten, ist innerhalb einer grossen Bandbreite durchschnittlich kontaktfähig und ist überdurchschnittlich verantwortungsbewusst. Auf Grund der in einer Selektionssituation durchgeführten Testung könnte es sich bei diesem Bild jedoch auch um das Ergebnis der Einstufung der sozialen Erwünschtheit dieser drei Dimensionen handeln. Die definitive Antwort auf die Frage nach der Ursache verschiedenen Score-Verteilungen in den drei Skalen lässt sich schlussendlich nur in einer experimentellen Studie bestimmen.

Tabelle 6.27

Verteilungskennwerte der drei Skalen des Leadership-Fragebogens

	Minimum	Maximum	*M*	*SD*	Schiefe	Exzess
Durchsetzungsfähigkeit	10	39	23.43	3.55	.13	.35
Kontaktfähigkeit	10	40	26.30	5.14	-.27	-.27
Verantwortungsbewusstsein	10	38	27.45	4.31	-.72	.81

Anmerkung. *N* = 7'871.

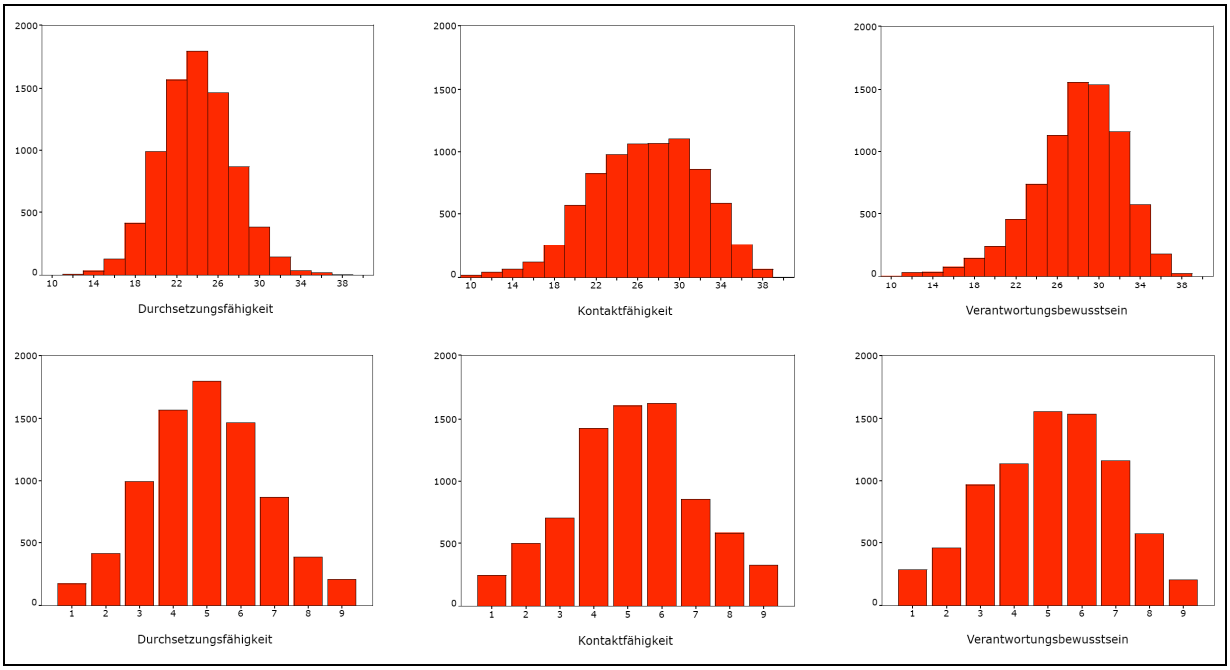


Abbildung 6.9 Rohwert- und Stanine-Verteilungen der drei Skalen der 30-Item-Version des Leadership-Fragebogens.

Ich möchte an dieser Stelle noch auf eine Möglichkeit der Auswertung des Leadership-Fragebogens eingehen, welche sich auf Grund des Einsatzes des Wertequadrates ergibt und so deutlich von derjenigen bei herkömmlichen Persönlichkeits-Inventaren abweicht. Üblicherweise transformiert der Diagnostiker bei der Auswertung des Fragebogen-Resultates den Rohwert mit Hilfe der auf den Kandidaten passenden Tabelle in einen Normwert. Somit lässt sich das individuelle Ergebnis mit denjenigen anderer Kandidaten vergleichen. Ein auf dem Wertequadrat basierender Persönlichkeits-Fragebogen lässt sich nun aber – den Ansatz von Gloor (1993) weiterverfolgend – auch bezogen auf die einzelnen Wertequadranten auswerten. Dabei gebe ich für jeden davon einzeln an, wie häufig der Kandidat die entsprechende Verhaltensalternative gewählt hat und vergleiche seinen Wert mit einem Durchschnittswert. In Abbildung 6.10 habe ich einen konkreten Fall aus einem Pilotversuch mit Offiziers-Aspiranten dargestellt.

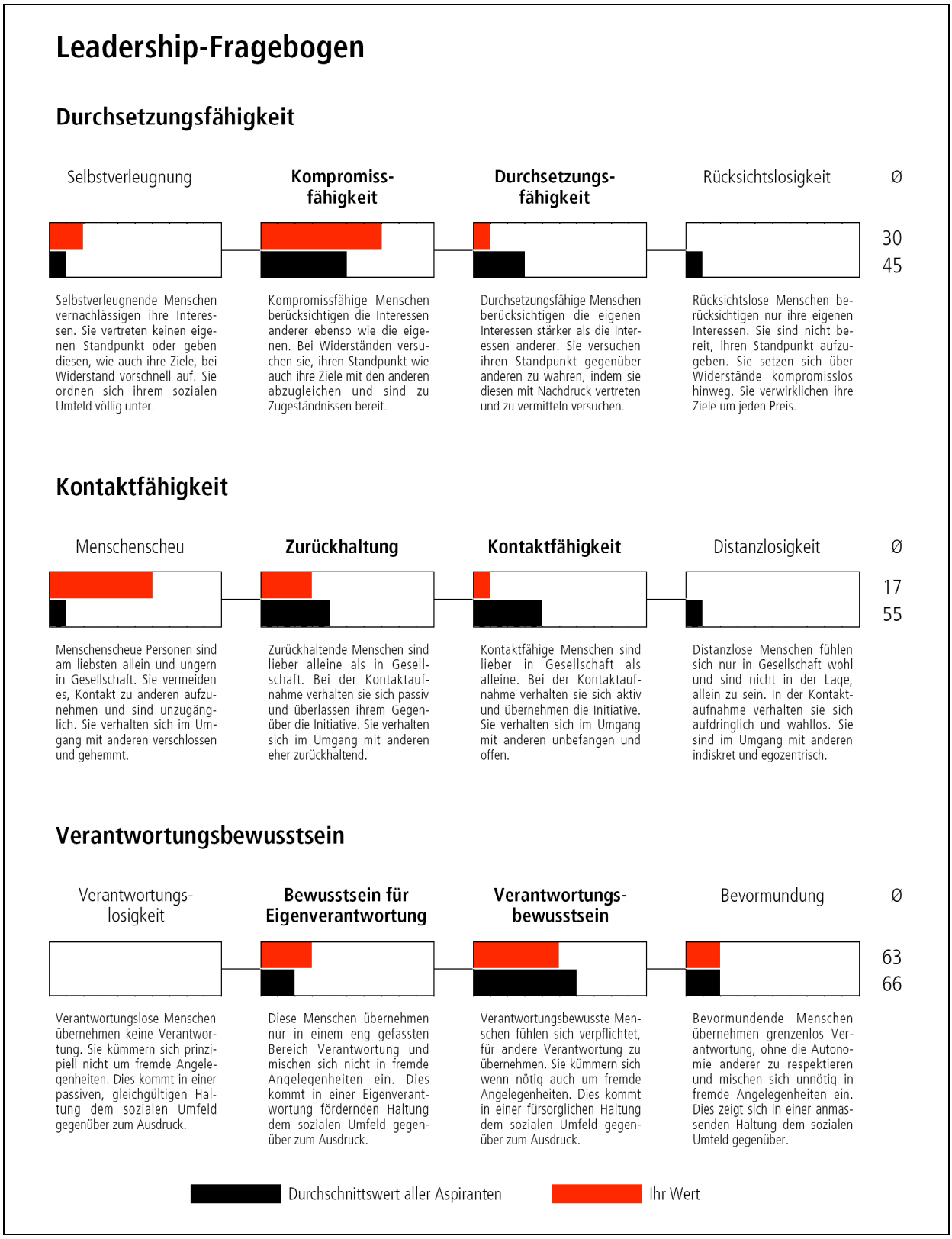


Abbildung 6.10 Wertequadrat-basierte Auswertung des Leadership-Fragebogens.

Diese Art der Auswertung bietet mehrere Vorteile: Für den Kandidaten sind die Ergebnisse sehr anschaulich und leicht verständlich dargestellt. Dieser – und natürlich auch der Diagnostiker – ist so in der Lage, viele Einzelinformationen auf einen Blick zu erfassen und genau nachvollziehen zu können, wie der ebenfalls aufgeführte Durchschnittswert entstanden ist. Dieser Aspekt ist gerade für den Diagnostiker bei der Interpretation und der Rückmeldung der Ergebnisse sehr wichtig und bietet ihm vielfältige Ansatz- und Erklärungsmöglichkeiten. Hier zeigt sich dann der von Birkhan (2007) beschriebene grosse Vorteil der Wertequadrat-Methode, indem der Diagnostiker sehr anschaulich, beleg- und nachvollziehbar konstruktive Rückmeldungen geben kann, welche Möglichkeiten eröffnen und Defizite als Chance für Entwicklungen aufzeigen.

Dieses Kapitel kurz zusammenfassend, kann ich sagen, dass es uns gelungen ist, die drei Dimensionen des Leadership-Fragebogens zu operationalisieren, so dass ich sie faktorenanalytisch bestätigen konnte. Ganz zufriedenstellend ist der Fragebogen jedoch nicht ausgefallen: Die Skala Durchsetzungsfähigkeit ist zu wenig reliabel, die drei Faktoren klären insgesamt zu wenig Varianz auf und die Korrelation zwischen den Skalen Kontaktfähigkeit und Verantwortungsbewusstsein fällt tendenziell zu hoch aus. Im nachfolgenden Kapitel stelle ich Resultate der an einem umfangreichen Datensatz durchgeführten Überprüfungsstudie vor und suche nach Möglichkeiten, eben erwähnte Schwachpunkte zu beheben. Weiter gehe ich der Frage nach, ob es uns auch gelungen ist, ein für die Stellungspflichtigen attraktives, das heisst gut akzeptiertes Instrument zu entwickeln und untersuche abschliessend den Zusammenhang zu Intelligenzmassen und anderen Persönlichkeitsskalen.

6.5 Literaturverzeichnis

Ajzen, I. (2002). Residual effects of past on later behavior: Habituation and reasoned action perspectives. *Personality and Social Psychology Review, 6*, 107–122.

Albanese, R., & Van Fleet, D. D. (1985). Rational behavior in groups: The free-riding tendency. *Academy of Management Review, 10*, 244–255.

Anderson, L., & Wilson, S. (1997). Critical incident technique. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measure methods in industrial psychology* (pp. 89–112). Palo Alto, CA: Davies-Black.

Angleitner, A., Buss, D. A., & Demtröder, A. I. (1990). A cross-cultural comparison using the Act Frequency Approach (AFA) in West Germany and the United States. *European Journal of Personality, 4*, 187–207.

Annen, H. (2000). *Förderwirksame Beurteilung. Aktionsforschung in der* Schweizer Armee. Frauenfeld: Huber.

Bauhofer, R., Bösser, D. & Henggeler, C. (2003). *Leadership-Fragebogen zur Kaderselektion bei der Rekrutierung: Erste Validierung und motivationale Einflüsse.* Unveröff. Bericht, Universität Zürich.

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6*, 205–212.

Birkhan, G. (2007). Das unipolare und das bipolare Eigenschaftsmodell in Diagnostik und Beratung. In F. Westermann (Hrsg.), *Entwicklungsquadrat. Theoretische Fundierung und praktische Anwendungen* (S. 21–29). Göttingen: Hogrefe.

Boss, P. (2005). Assessment in der Arbeitswelt – Kriterien für eine bewerberzentrierte Personalauswahl. In M. Rehbinder (Hrsg.), *Psychologische Aspekte im Recht der Personalführung* (S. 21–45). Bern: Stämpfli.

Brannick, M. T., & Levine, E. L. (2002). *Job analysis. Methods, research, and applications for human resource management in the new millennium.* Thousand Oaks, CA: Sage.

Bühler Ruedin, A. & Selk, T. (2001). *Konzeption eines Fragebogens zur Erfassung von Führungspotenzial.* Unveröff. Bericht, Universität Zürich.

Buss, D. M., & Craik, K. H. (1980). The frequency concept of disposition: Dominance and prototypically dominant acts. *Journal of Personality, 48,* 379–392.

Buss, D. M., & Craik, K. H. (1983). The Act Frequency Approach to personality. *Psychological Review, 90,* 105–126.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Cascio, W. F., & Aguinis, H. (2005). *Applied psychology in human resource management* (6th ed.). Upper Saddle River, NJ: Pearson.

Conley, P. R., & Sackett, P. R. (1987). Effects of using high- versus low-performing job incumbents as sources of job-analysis information. *Journal of Applied Psychology, 72*, 434–437.

Cornelius, E. T., DeNisi, A. S., & Blencoe, A. G. (1984). Expert and naïve raters using the PAQ: Does it matter? *Personnel Psychology, 37*, 453–464.

Deiss, E., Emerson, A., Imper, A. & Maier, V. (2002). *Überprüfung und Weiterentwicklung des Leadership-Fragebogens zur Erfassung von Führungspotenzial.* Unveröff. Bericht, Universität Zürich.

Dierdorff, E. C., & Morgeson, F. P. (2007). Consensus in work role requirements: The influence of discrete occupational context on role expectations. *Journal of Applied Psychology, 92*, 1228–1241.

Dierdorff, E. C., & Wilson, M. A. (2003). A meta-analysis of job analysis reliability. *Journal of Applied Psychology, 88*, 635–646.

DIN (2002). *DIN 33430: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen*. Berlin: Beuth.

Eckardt, H. H. & Schuler, H. (1992). Berufseignungsdiagnostik. In R. S. Jäger & F. Petermann (Hrsg.), *Psychologische Diagnostik* (2. Aufl., S. 533–551). Weinheim: Psychologie Verlags Union.

Etzel, S. (1999). *Multimediale, computergestützte diagnostische Verfahren: Neue Perspektiven für die Managementdiagnostik.* Aachen: Shaker.

Evers, A. (2001). The revised Dutch rating system for test quality. *International Journal of Testing, 1*, 155–182.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327–358.

Fuhrer, G. (1985). *Das Anforderungsprofil militärischer Chefs.* Unveröff. Diplomarbeit. Zürich: ETHZ, Abteilung für Militärwissenschaften.

Gael, S. (1983). *Job analysis: A guide to assessing work activities.* San Francisco, CA: Jossey-Bass.

Gael, S. (Ed.). (1988). *The job analysis handbook for business, industry, and government.* New York, NY: Wiley.

Gatewood, R. D., Field, H. S., & Barrick, M. (2008). *Human resource selection* (6th ed.). Mason, OH: Thomson South-Western.

Gloor, A. (1993). *Die AC-Methode. Assessment Center. Führungskräfte beurteilen und fördern.* Zürich: Orell Füssli.

Gloor, A. (2007). *Das Werte- und Entwicklungsquadrat.* Unterlagen zur Lehrveranstaltung „Das Wertequadrat als Denkmuster im HRM". Heruntergeladen am 29. Juni 2010 von ftp://ftp.unizh.ch/hrm/03_studium/veranstaltungen/hrm_2/Folien_HRMII_SS07/2_Gastreferat_Armin_Gloor.pdf

Gonin, N. O. (1993). *Unteroffiziersselektion – Eine Untersuchung am Beispiel der Schweizer Armee.* Frauenfeld: Huber.

Green, S. B., & Stutzman, T. (1986). An evaluation of methods to select respondents to structured job-analysis questionnaires. *Personnel Psychology, 39*, 543–564.

Green, S. B., & Veres, J. G. (1990). Evaluation of an index to detect inaccurate respondents to a task analysis inventory. *Journal of Business and Psychology, 5*, 47–61.

Guion, R. M. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahwah, NJ: Erlbaum.

Haitovsky, Y. (1969). Multicollinearity in regression analysis: A comment. *Review of Economics and Statistics, 51*, 486–489.

Harvey, R. J. (1991). Job analysis. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology: Vol. 2* (2nd ed., pp. 71–163). Palo Alto, CA: Consulting Psychologists Press.

Hell, B., Ptok, C. & Schuler, H. (2007). Methodik zur Ermittlung und Validierung von Anforderungen an Studierende (MEVAS). *Zeitschrift für Arbeits- und Organisationspsychologie, 51*, 88–95.

Helwig, P. (1948). Das Wertequadrat. *Psyche, 2*, 121–127.

Hoenle, S. (1996). *Führungskultur in der Schweizer Armee.* Frauenfeld: Huber.

Höft, S. & Schuler, H. (2005). Empirische Arbeits- und Anforderungsanalysen: Ein Anwendungsbeispiel mit einem kombinierten aufgaben-, verhaltens- und eigenschaftsorientierten Analyseansatz. In K. Sünderhauf, S.

Stumpf & S. Höft (Hrsg.), *Assessment Center. Von der Auftragsklärung bis zur Qualitätssicherung. Ein Handbuch von Praktikern für Praktiker* (S. 72–88). Lengerich: Pabst.

Hossiep, R. & Paschen, M. (2003). *Das Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung* (2. vollst. überarb. Aufl.). Göttingen: Hogrefe.

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Klehe, U.-C. (2004). The impact of job complexity and study design on situational and behavior description interview validity. *International Journal of Selection and Assessment, 12*, 262–273.

Huffcutt, A. I., Weekley, J. A., Wiesner, W. H., Degroot, T. G., & Jones, C. (2001). Comparison of situational and behavior description interview questions for higher-level positions. *Personnel Psychology, 54*, 619–644.

Hughes, G. L., & Prien, E. P. (1989). Evaluation of task and job skill linkage judgments used to develop test specifications. *Personnel Psychology, 42*, 283–292.

Imper, A. & Maier, V. (2003). *Validierung eines Fragebogens zur Erfassung von Führungspotenzial bei Stellungspflichtigen.* Unveröff. Bericht, Universität Zürich.

Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. *Journal of Applied Psychology, 67*, 577–580.

Janz, T. (1989). The patterned behavior description interview: The best prophet of the future is the past. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 158–168). Newbury Park, CA: Sage.

Jetter, W. (2008). *Effiziente Personalauswahl. Durch strukturierte Einstellungsgespräche die richtigen Mitarbeiter finden* (3., aktualisierte überarb. u. erw. Aufl.). Stuttgart: Schäffer-Poeschel.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*, 31–36.

Kannheiser, W. (1995). Erfassung der Anforderungen einer konkreten Position. In W. Sarges (Hrsg.), *Management-Diagnostik* (2. vollst. überarb. u. erw. Aufl., S. 141–148). Göttingen: Hogrefe.

Kanning, U. P. (2002). Nicht-standardisierte Methoden. In U. P. Kanning & H. Holling (Hrsg.), *Handbuch personaldiagnostischer Instrumente* (S.

118–124). Göttingen: Hogrefe.

Kanning, U. P. (2004). *Standards der Personaldiagnostik.* Göttingen: Hogrefe.

Kanning, U. P. & Holling, H. (Hrsg.). (2002). *Handbuch personaldiagnostischer Instrumente.* Göttingen: Hogrefe.

Kidwell, R. E., & Bennett, N. (1993). Employee propensity to withhold effort: A conceptual model to intersect three avenues of research. *Academy of Management Review, 18*, 429–456.

Kline, P. (1998). *The new psychometrics. Science, psychology and measurement.* London: Routledge.

Koch, A., Kici, G., Strobel, A. & Westhoff, K. (2006). Anforderungsanalysen nach DIN 33430: exemplarisch für die Position eines Dozenten im Arbeitsschutz. In K. Westhoff (Hrsg.), *Nutzen der DIN 33430. Praxisbeispiele und Checklisten* (S. 85–93). Lengerich: Pabst.

Krampen, G. (1991). *Fragebogen zu Kompetenz- und Kontrollüberzeugungen (FKK). Handanweisung.* Göttingen: Hogrefe.

Krüger, C. & Amelang, M. (1995). Bereitschaft zu riskantem Verhalten als Trait-Konstrukt und Test-Konzept: Zur Entwicklung eines Fragebogens auf der Basis des Handlungs-Häufigkeits-Ansatzes. *Diagnostica, 41*, 35–52.

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 65*, 422–427.

Li, W.-D., Wang, Y.-L., Taylor, P., Shi, K., & He, D. (2008). The influence of organizational culture on work-related personality requirement ratings: A multilevel analysis. *International Journal of Selection and Assessment, 16*, 366–384.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Psychologie Verlags Union.

Lievens, F., & Sanchez, J. I. (2007). Can training improve the quality of inferences made by raters in competency modeling? A quasi-experiment. *Journal of Applied Psychology, 92*, 812–819.

Lievens, F., Sanchez, J. I., Bartram, D., & Brown, A. (2010). Lack of consensus among competency ratings of the same occupation: Noise or substance? *Journal of Applied Psychology, 95*, 562–571.

Lievens, F., Sanchez, J. I., & De Corte, W. (2004). Easing the inferential leap in competency modeling: The effects of task-related information and sub-

ject matter expertise. *Personnel Psychology, 57*, 881–904.

Lindley, P., Bartram, D., & Kennedy, N. (2008). *EFPA review model for the description and evaluation of psychological tests. Test review form and notes for reviewers* (Version 3.42). European Federation of Psychologists' Associations. Heruntergeladen am 1. Juli 2010 von www.efpa.eu/download/9044bd41c7953b956876e06c797f8c9f

Marx, W. & Läge, D. (1995). *Der ideologische Ring.* Göttingen: Hogrefe.

Maurer, T. J., & Tross, S. A. (2000). SME committee vs. field job analysis rating: Convergence, cautions, and a call. *Journal of Business and Psychology, 14*, 489–499.

Mayring, P. (2008). *Qualitative Inhaltsanalyse. Grundlagen und Techniken* (10. neu ausgestattete Aufl.). Weinheim: Beltz.

McCormick, E. J. (1979). *Job analysis: Methods and applications*. New York, NY: AMACOM.

Moosbrugger, H. & Rauch, W. (2005). Klassische Testtheorie. In K. Westhoff, L. J. Hellfritsch, L. F. Hornke, K. D. Kubinger, F. Lang, H. Moosbrugger, A. Püschel & G. Reimann (Hrsg.), *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (2. überarb. Aufl., S. 182–186). Lengerich: Pabst.

Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology, 82*, 627–655.

Moroge, F. & Schibli, M. (2002). *Weiterentwicklung eines Fragebogens zur Beurteilung von Führungskompetenz bei Stellungspflichtigen.* Unveröff. Bericht, Universität Zürich.

Motowidlo, S. J. (1999). Asking about past behavior versus hypothetical behavior. In R. W. Eder & M. M. Harris (Eds.), *The employment interview handbook* (pp. 179–190). Thousand Oaks, CA: Sage.

Motowidlo, S. J., Carter, G. W., Dunnette, M. D., Tippins, N., Werner, S., Burnett, J. R., et al. (1992). Studies of the structural behavioral interview. *Journal of Applied Psychology, 77*, 571–587.

Netemeyer, R. G., Bearden, W. O., & Sharma, S. (2003). *Scaling procedures. Issues and applications.* Thousand Oaks, CA: Sage.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory.* New York, NY: McGraw-Hill.

Ostendorf, F. & Angleitner, A. (2004). *NEO-PI-R. NEO-Persönlichkeitsinventar nach Costa und McCrae. Revidierte Fassung. Manual.* Göttingen: Hogrefe.

Ouellette, J. A., & Wood, W. (1998). Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological Bulletin, 124*, 54–74.

Pulakos, E. D., & Schmitt, N. (1995). Experience-based and situational interview questions: Studies of validity. *Personnel Psychology, 48*, 289–308.

Reimann, G. (2005). Arbeits- und Anforderungsanalyse. In K. Westhoff, L. J. Hellfritsch, L. F. Hornke, K. D. Kubinger, F. Lang, H. Moosbrugger, A. Püschel & G. Reimann (Hrsg.), *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430* (2. überarb. Aufl., S. 111–127). Lengerich: Pabst.

Rockwell, R. C. (1975). Assessment of multicollinearity: The Haitovsky test of the determinant. *Sociological Methods and Research, 3,* 308–320.

Rose, D. S., & Baydoun, R. (1995). Content validity: A neglected strategy for developing managerial selection tests? *Current Psychology: Research & Reviews, 14*, 138–151.

Rückert, J. (1993). *Psychometrische Grundlagen der Diagnostik.* Göttingen: Hogrefe.

Ryan, A. M., & Ployhart, R. E. (2000). Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *Journal of Management, 26*, 565–606.

Sanchez, J. L. & Fraser, S. L. (1992). On the choice of scales for task analysis. *Journal of Applied Psychology, 77*, 545–553.

Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*, 735–746.

Schleicher, D. J., Venkataramani, V., Morgeson, F. P., & Campion, M. A. (2006). So you didn't get the job … Now what do you think? Examining opportunity-to-perform fairness perceptions. *Personnel Psychology, 59*, 559–590.

Schuler, H. (2002). *Das Einstellungsinterview*. Göttingen: Hogrefe.

Schuler, H., Prochaska, M. & Frintrup, A. (2001). *Leistungsmotivationsinventar. Dimensionen berufsbezogener Leistungsorientierung. Manual.* Göttin-

gen: Hogrefe.

Schulz von Thun, F. (1989). *Miteinander Reden 2. Stile, Werte und Persönlichkeitsentwicklung.* Reinbek bei Hamburg: Rowohlt.

Schweizer Armee (1995). *Dienstreglement DR 95. Reglement 51.2 d.* Bern: EDMZ.

Schweizer Armee (1997). *Weisungen über Qualifikationen und Vorschläge zur Weiterausbildung (WQV 97). Reglement 51.13 d.* Bern: EDMZ.

Society for Industrial and Organizational Psychology, Inc. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Bowling Green, OH: Author.

Stadelmann, J. (1998). *Führung unter Belastung. Ausgewählte Aspekte der Militärpsychologie.* Frauenfeld: Huber.

Steiger, R. (1999). *Menschenorientierte Führung. Anregungen für zivile und militärische Führungskräfte* (11. Aufl.). Huber: Frauenfeld.

Steiger, R. & Annen, H. (1997). „ACABO" – Das Assessment Center als Selektionsinstrument für angehende Berufsoffiziere. *Allgemeine Schweizerische Militär-Zeitschrift, 2*, 9–11.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Hillsdale, NJ: Erlbaum.

Tannenbaum, R. J., & Wesley, S. (1993). Agreement between committee-based and field-based job analyses: A study in the context of licensure testing. *Journal of Applied Psychology, 78*, 975–980.

Taylor, P. J., & Small, B. (2002). Asking applicants what they would do versus what they did do: A meta-analytic comparison of situational and past behaviour employment interview questions. *Journal of Occupational and Organizational Psychology, 75*, 277–294.

Voskuijl, O. F., & van Sliedregt, T. (2002). Determinants of interrater reliability of job analysis: A meta-analysis. *European Journal of Psychological Assessment, 18*, 52–62.

Weldon, E. & Gargano, G. M. (1985). Cognitive effort in additive task groups: The effects of shared responsibility on the quality of multiattribute judgments. *Organizational Behavior and Human Decision Processes, 36*, 348–361.

Weekley, J. A., Ployhart, R. E., & Holtz, B. C. (2006). On the development of situational judgment tests: Issues in item development, scaling, and

scoring. In J. A. Weekley & R. R. Ployhart (Eds.), *Situational judgment tests: Theory, management, and application* (pp. 157–182). Mahwah, NJ: Erlbaum.

Westermann, F. (Hrsg.). (2007a). Entwicklungsquadrat. Theoretische Fundierung und praktische Anwendungen. Göttingen: Hogrefe.

Westermann, F. (2007b). Wer einen Schlüssel hat, der Türen öffnet, braucht nicht durch die Wand zu gehen! Das Entwicklungsquadrat – eine Einführung. In F. Westermann (Hrsg.), *Entwicklungsquadrat. Theoretische Fundierung und praktische Anwendungen* (S. 9–19). Göttingen: Hogrefe.

Wexley, K. N., & Silverman, S. B. (1978). An examination of differences between managerial effectiveness and response patterns on a structured job analysis questionnaire. *Journal of Applied Psychology, 63*, 646–649.

Wheaton, G. R., & Whetzel, D. L. (2007). Contexts for developing applied measure instruments. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement. Industrial psychology in human resources management* (pp. 1–11). Mahwah, NJ: Erlbaum.

Wilson, M. A., Harvey, R. J., & Macy, B. A. (1990). Repeating items to estimate the test-retest reliability of task inventory ratings. *Journal of Applied Psychology, 75*, 158–163.

Wirtz, M. & Caspar, F. (2002). *Beurteilerübereinstimmung und Beurteilerreliabilität. Methoden zur Bestimmung und Verbesserung der Zuverlässigkeit von Einschätzungen mittels Kategoriensystemen und Ratingskalen.* Göttingen: Hogrefe.

Wohlers, A. J., & London, M. (1989). Ratings of managerial characteristics: Evaluation difficulty, co-worker agreement, and self-awareness. *Personnel Psychology, 42*, 235–261.

Wrzesniewski, A., & Dutton, J. E. (2001). Crafting a job: Revisioning employees as active crafters of their work. *Academy of Management Review, 26*, 179–201.

Zollinger, P. (1997). Das Anforderungsprofil für den Berufsoffizier der Schweizer Armee. *Allgemeine Schweizerische Militär-Zeitschrift, 2*, 5–7.

Anhang 6.1 Anleitung für die Erstellung des Kategoriensystems

Als Ausgangsmaterial dienen Kärtchen mit jobrelevanten Verhaltensweisen für die Position eines Unteroffiziers, welche anhand von Interviews mit Berufsoffizieren generiert wurden. Ziel der Kategorisierung ist die Bildung der Dimensionen oder Kategorien des Anforderungsprofils. Es geht nicht darum, dass jede Verhaltensweise genau der richtigen Kategorie zugeordnet wird. Wichtig ist nur die Bestimmung der Kategorien.

In einem ersten Schritt (erste Reduktion der Komplexität) sollen die Kärtchen eines nach dem anderen einem sich laufend erweiternden und verändernden Kategoriensystem zugeteilt werden. Dabei sollen nicht zu viele Kategorien gebildet werden (ca. 10), damit der Überblick behalten werden kann. Die (vorläufigen) Kategoriennamen werden auf ein A4-Blatt geschrieben und die entsprechenden Kärtchen darauf gelegt. Nicht zuordenbare Kärtchen legt man auf einen separaten Stapel (diesen möglichst klein halten!).

Beispiel:	1. Kärtchen: *Küche sauber halten*	Bildung der Kategorie „Haushalt" (auf A4-Papier notieren und Kärtchen dazu legen)
	2. Kärtchen: *Fenster reinigen*	Kärtchen zum A4-Blatt „Haushalt" legen
	3. Kärtchen: *Kinder zur Schule bringen*	Bildung der Kategorie „Erziehung" (auf A4-Papier notieren und Kärtchen dazu legen)
	4. Kärtchen: *Zähneputzen kontrollieren*	Kärtchen zum A4-Blatt „Erziehung" legen
	5. Kärtchen: *zum Frisör gehen*	Bildung der Kategorie „Schönheitspflege" (auf A4-Papier notieren und Kärtchen dazu legen)
	usw.	

Es kann vorkommen, dass im Verlaufe des Sortierens Kategorien umbenannt oder aufgeteilt werden müssen.

In einem zweiten Schritt (zweite Reduktion der Komplexität) nimmt man sich die einzelnen Kategorien nochmals vor und sortiert die dazu geordneten Kärtchen zu Unterkategorien (ca. drei pro (Ober-) Kategorie). Nicht passende Kärtchen werden anderen (Ober-) Kategorien zugeordnet oder auf den Stapel „nicht zuordenbar" gelegt. Kärtchenstapel pro Unterkategorie mit Büroklammern „fixieren" und auf einem leeren Kärtchen den Unterkategorienamen notieren und dieses oben auf den Stapel legen.

Beispiel (Ober-) Kategorie „Haushalt":

1. Kärtchen: *Küche sauber halten*	Bildung der Unterkategorie „Reinigung" (auf leeres Kärtchen notieren und Kärtchen dazu legen)
2. Kärtchen: *Fenster reinigen*	Kärtchen zum Kärtchenstapel „Reinigung" legen
3. Kärtchen: *Topfpflanzen giessen*	Bildung der Unterkategorie „Unterhalt der Wohnung" (auf leeres Kärtchen notieren und Kärtchen dazu legen)
4. Kärtchen: Kleider waschen	Bildung der Unterkategorie „Waschen, Bügeln" (auf leeres Kärtchen notieren und Kärtchen dazu legen) Unterkategorie „Reinigung" in „Reinigung der Wohnung" umbenennen.
5. Kärtchen: *Kochen*	Bildung der Unterkategorie „Kochen" (auf leeres Kärtchen notieren und Kärtchen dazu legen)
usw.	

In einem dritten Schritt wird versucht, die Kärtchen vom Stapel „nicht zuordenbar" doch noch irgendwo zuzuteilen, evtl. in einer neuen Unterkategorie. Der unzuordenbare Rest beiseite legen.

In einem vierten Schritt dienen die Unterkategorie-Kärtchenstapel als Ausgangsmaterial für die Bildung des definitiven Kategoriensystems. Das Vorgehen ist dem in Schritt 1 ähnlich. Es ist nun aber darauf zu achten, dass sich die (Ober-) Kategorien gut voneinander unterscheiden und keine Überlappungen aufweisen. Es sollen ca. zehn (Ober-) Kategorien gebildet werden. (Ein Anforderungsprofil mit 20 Dimensionen ist für die praktische Verwendung ungeeignet.) Die Namen dieser (Ober-) Kategorien müssen sich nicht mit den in Schritt 1 gewählten decken, sondern bilden den Überbegriff der dieser Kategorie zugeteilten Unterkategorien. Achtung: Die (Ober-) Kategorien nicht zu allgemein wählen (wie z.B. Selbst-, Sozial-, Führungs- und Methodenkompetenz und Fachwissen). Die Kärtchenstapel mit den Unterkategorien steckt man pro Kategorie in ein Couvert, welches den Namen der (Ober-) Kategorie trägt. Zur Sicherheit das Kategorisierungssystem auf einem Blatt aufzeichnen:

Haushalt
- Reinigung der Wohnung
- Waschen, Bügeln
- Kochen

usw.

Erziehung
- Kontrolle der Körperpflege
- Kontrolle der Schulaufgaben
- Durchsetzen von Verhaltensnormen

Anhang 6.2 Anforderungsprofil für Gruppenführer

Analysefähigkeit	macht eine angepasste Lagebeurteilung; sucht und findet eine Lösung; kann die Konsequenzen ein- und abschätzen; besitzt eine schnelle Auffassungsgabe; begreift neue Sachen schnell und kann sie umsetzen
Organisations- und Planungsfähigkeit	denkt, plant und handelt vorausschauend; hat und behält den Überblick; setzt Prioritäten; nimmt sich genügend Zeit für die Planung und Vorbereitung der Ausbildungsinhalte; delegiert und erteilt Aufträge
physische & psychische Belastbarkeit	kann physisch mithalten, ist leistungsfähig, körperlich fit; ist psychisch belastbar; handelt ruhig und überlegt; zeigt eine konstante Arbeitsleistung; bleibt auch unter Druck sachlich, ausgeglichen und zuverlässig
Leistungsbereitschaft & Engagement	will eine gute Leistung erzielen; ist motiviert, zeigt Einsatz und packt selbst mit an; hat Ziele, die er erreichen will und behält die Zielerreichung im Auge; ist selbst von der Sache überzeugt und begeistert und motiviert so seine Leute; gibt bei Rückschlägen nicht auf
Gewissenhaftigkeit	ist zuverlässig und besitzt Pflichtbewusstsein; ist sorgfältig, diszipliniert, zeitgerecht und auftragstreu; ist integer und hält sich an Regeln; befolgt, besitzt und vertritt gewisse Werte und Normen; ist loyal dem Chef und seiner Gruppe gegenüber
Offenheit & Flexibilität	ist offen für Neues und für Verbesserungsvorschläge; sucht nach Alternativen, ist innovativ, kreativ und hat Ideen; reagiert flexibel und anpassungsfähig; denkt positiv und hat eine positive Grundeinstellung; akzeptiert andere Lösungsvorschläge und Meinungen
Selbstreflexion	hinterfragt sein Handeln; ist selbstkritisch und schätzt sich realistisch ein; ist ehrlich sich selbst und anderen gegenüber; kennt seine Schwächen und kann diese auf eine gute Art kompensieren; ist sich seiner Stärken bewusst
Teamfähigkeit	ist kontaktfreudig, geht auf Menschen zu; integriert sich in eine Gruppe; kann mit unterschiedlichsten Personen angemessen umgehen; stellt Gruppenkohäsion her und integriert alle ins Team; bietet seine Hilfe an
Fürsorglichkeit	ist fürsorglich gegenüber seiner Gruppe; behandelt seine Unterstellten und die Mitmenschen respektvoll; nimmt sich Zeit für die Unterstellten; kennt seine Unterstellten; schaut zuerst für seine Leute, stellt sich und seine Bedürfnisse in den Hintergrund
Kommunikationsfähigkeit	kommuniziert offen, direkt und ehrlich; kann angemessen argumentieren und artikulieren; bleibt bei Meinungsverschiedenheiten ruhig und korrekt; kommuniziert sachlich und in einem der Situation angemessenen Tonfall; erklärt den Sinn und Zweck einer Aufgabe oder einer Übung verständlich
Durchsetzungsfähigkeit	kann sich durchsetzen; entscheidet auch gegen Widerstände; handelt konsequent und zielorientiert; steht für seine Meinung ein, bringt eigenen Standpunkt ein; kann sich Gehör verschaffen
Verantwortungsübernahme	übernimmt die Verantwortung, besitzt Verantwortungsbewusstsein; ist ein Vorbild in der Erscheinung und seinen Handlungen; sanktioniert ungebührendes Verhalten der Unterstellten; greift bei Falschverhalten korrigierend ein; führt und behält den Führungsanspruch
Selbstsicherheit	besitzt Selbstvertrauen, ist selbstbewusst, weiss, dass er etwas kann; ist ein Leadertyp; strahlt eine gewisse Autorität und Selbstsicherheit aus; hat ein sicheres Auftreten; stellt sich dem Konflikt und ist kritikfähig

Anhang 6.3 Anforderungsprofil für Zugführer

Analysefähigkeit	besitzt eine gute Analysefähigkeit und ein strukturiertes Denkvermögen; verfügt über eine hohe Konzentrationsfähigkeit; kann vernetzt denken und erkennt Zusammenhänge; führt eine Problem-/Situationsanalyse durch; kann aus einer Fülle von Informationen die wichtigen erkennen
Organisations- & Planungsfähigkeit	sorgt für eine gute Planung und Organisation; geht systematisch und strukturiert vor; leitet aus einer Situation das richtige Handeln ab; verschafft sich und hat den Überblick; hat die Handlungsabläufe im Griff
Allgemeinbildung	hat Lehre oder Matura abgeschlossen; hat eine gute Allgemeinbildung; verfügt über einen guten und korrekten schriftlichen Ausdruck
physische & psychische Belastbarkeit	ist psychisch und physisch belastbar; ist stressresistent; ist zäh, beisst sich durch; bleibt in hektischen Situationen gelassen, handelt angemessen und erbringt eine gute Leistung; besitzt Frustrationstoleranz
Leistungsbereitschaft & Engagement	verfügt über einen ausgeprägten Leistungswillen; zeigt Engagement; identifiziert sich mit seiner Aufgabe; ist beharrlich und zeigt Durchhaltewillen; hat den Willen zum Erfolg; sieht in der Schwierigkeit eine Herausforderung; ist selbständig und zeigt Eigeninitiative
Gewissenhaftigkeit & Loyalität	arbeitet genau und gewissenhaft; ist auftragstreu; ist verantwortungs- & pflichtbewusst; ist loyal gegenüber seinen Vorgesetzten & Unterstellten; ist integer; nimmt sich Zeit für eine seriöse Vorbereitung oder Abklärung
Offenheit & Flexibilität	ist offen für andere Kulturen, Sprachen und Ansichten; hat den Mut, neue Ideen zu generieren und auszuprobieren; ist flexibel in seiner Denkweise und seinen Handlungen; ist anpassungsfähig; ist ein Optimist
Selbstreflexion	besitzt Selbstreflexion; ist selbstkritisch; reflektiert und hinterfragt sein Handeln; lernt aus seinen Fehlern; bleibt realistisch; ist bescheiden; kennt seine Schwächen, steht dazu; weiss, was sein Verhalten bewirkt
Kooperationsfähigkeit	arbeitet für die Auftragserfüllung mit anderen zusammen; kann sich ein- und unterordnen; ist kameradschaftlich; ist hilfsbereit; spricht sich mit seinen Kameraden und Unterstellten ab, um ans Ziel zu kommen
Einfühlungsvermögen	verfügt über Einfühlungsvermögen; hat ein Gespür für seine Mitmenschen; hat Verständnis für seine Unterstellten; zeigt Fürsorge für seine Leute; wertschätzt seine Unterstellten und behandelt sie respektvoll
Konflikt- & Kritikfähigkeit	kann mit Konflikten umgehen, erkennt sie, geht sie offen an, kann sie lösen; spricht Konflikte direkt an; fragt nach bei Unstimmigkeiten oder bei Problemen; sucht und findet einen Konsens; kann mit Kritik umgehen; reagiert differenziert und sachlich, nimmt nicht alles persönlich
Kommunikationsfähigkeit	ist kommunikativ; informiert umfassend, offen und ehrlich; sucht das Gespräch; hört aktiv zu; Kommuniziert klar und deutlich; Kommuniziert sachlich und in einem der Situation angemessenen Tonfall
Durchsetzungsfähigkeit	kann sich durchsetzen; ist hartnäckig, damit das Ziel erreicht werden kann; erzwingt den Erfolg; ist konsequent; duldet unkorrektes Verhalten nicht; verfolgt die Konsequenzen; ist entscheidungsfreudig
Verantwortungsübernahme	übernimmt die Verantwortung, auch für das Handeln anderer; nimmt Einfluss; stellt sich auch unangenehmen Situationen; trägt die Konsequenzen; schützt seine Leute; greift bei Bedarf korrigierend ein
Auftreten als Chef	hat ein sicheres Auftreten; besitzt eine stolze Körperhaltung; übernimmt die Führung; ist natürlich und authentisch; ist eine gefestigte Persönlichkeit; besitzt ein gutes Stolz- und Ehrgefühl; strahlt Selbstvertrauen aus; vermittelt Sicherheit; ist ein Vorbild; hält den Zug zusammen und entwickelt Korpsgeist; besitzt einen gesunden Menschenverstand; repräsentiert die Armee in einer guten Art und Weise

Anhang 6.4 Liste der 112 Verhaltensweisen der Umfrage zur Erstellung des Basis-Anforderungsprofils

Analysefähigkeit	Macht eine angepasste Lagebeurteilung
	Sucht und findet eine Lösung
	Kann die Konsequenzen einschätzen und abschätzen
	Besitzt eine schnelle Auffassungsgabe
	Begreift neue Sachen schnell und kann sie umsetzen
	Verfügt über eine hohe Konzentrationsfähigkeit und ein gutes Gedächtnis
	Kann vernetzt denken und erkennt Zusammenhänge
	Kann aus einer Fülle von Informationen die wichtigen erkennen
Planungs- & Organisationsfähigkeit	Denkt, plant und handelt vorausschauend
	Hat und behält den Überblick
	Setzt Prioritäten
	Nimmt sich genügend Zeit für die Planung und Vorbereitung der Ausbildungsinhalte
	Delegiert und erteilt Aufträge
	Geht systematisch und strukturiert vor
	Kann die Konsequenzen seines Handelns abschätzen
	Setzt Prioritäten (Wichtigkeit und Dringlichkeit)
physische & psychische Belastbarkeit	Kann physisch mithalten, ist leistungsfähig, ist körperlich fit
	Ist psychisch belastbar
	Handelt ruhig und überlegt
	Zeigt eine konstante Arbeitsleistung
	Bleibt auch unter Druck sachlich, ausgeglichen und zuverlässig
	Ist bei Überraschungen nicht überfordert
	Hat seine Emotionen – positive wie negative – unter Kontrolle
	Besitzt Frustrationstoleranz
Leistungsbereitschaft & Engagement	Will eine gute Leistung erzielen
	Ist motiviert, zeigt Einsatz und packt selbst mit an
	Meldet sich freiwillig und ist bereit für Mehrarbeit
	Hat Ziele, die er erreichen will und behält die Zielerreichung im Auge
	Ist selbst von der Sache überzeugt und begeistert und motiviert so seine Leute
	Gibt bei Rückschlägen nicht auf
	Ist beharrlich und zeigt Durchhaltewillen
	Ist selbständig und zeigt Eigeninitiative
Gewissenhaftigkeit & Loyalität	Ist zuverlässig und besitzt Pflichtbewusstsein
	Ist sorgfältig, diszipliniert, zeitgerecht und auftragstreu
	Ist integer und hält sich an Regeln
	Befolgt, besitzt und vertritt gewisse Werte und Normen
	Macht auf kritische Punkte aufmerksam
	Ist loyal dem Chef und seinen Unterstellten gegenüber
	Ist verantwortungs- und pflichtbewusst
	Hinterfragt die an ihn gestellten Aufträge und überprüft den erhaltenen Befehl
Offenheit & Flexibilität	Ist offen für Neues und für Verbesserungsvorschläge
	Sucht nach Alternativen, ist innovativ, kreativ und hat Ideen
	Kann sich in verschiedene Situationen einfühlen
	Reagiert flexibel und anpassungsfähig
	Denkt positiv und hat eine positive Grundeinstellung
	Akzeptiert andere Lösungsvorschläge und Meinungen
	Ist bei der Zielerreichung flexibel
	Passt seinen Führungsstil bewusst an
Selbstreflexion	Hinterfragt sein Handeln
	Ist selbstkritisch und schätzt sich realistisch ein
	Ist ehrlich sich selbst und anderen gegenüber
	Kennt seine Schwächen und kann diese auf eine gute Art kompensieren
	Ist sich seiner Stärken bewusst
	Lernt aus seinen Fehlern und zieht den Mehrwert daraus
	Bleibt realistisch und ist bescheiden
	Weiss, was sein Verhalten bewirkt

Anhang 6.4 Liste der 112 Verhaltensweisen der Umfrage zur Erstellung des Basis-Anforderungsprofils (Fortsetzung)

Teamfähigkeit	Ist kontaktfreudig und geht auf Menschen zu Integriert sich in eine Gruppe Kann mit unterschiedlichsten Personen angemessen umgehen Stellt Gruppenkohäsion her und integriert alle ins Team Bietet seine Hilfe an Vermittelt weiter, wenn er selbst nicht helfen kann Kann sich ein- und unterordnen Kommuniziert mit seinen Kameraden um ans Ziel zu kommen
Fürsorglichkeit / Einfühlungsvermögen	Ist fürsorglich gegenüber seinen Unterstellten Behandelt seine Unterstellten und die Mitmenschen respektvoll Nimmt sich Zeit für die Unterstellten Kennt seine Unterstellten Kann sich in andere Menschen einfühlen Schaut zuerst für seine Leute, stellt sich und seine Bedürfnisse in den Hintergrund Verfügt über Einfühlungsvermögen Erkennt die Probleme seiner Leute, geht darauf ein und kann Lösungen anbieten
Konflikt- & Kritikfähigkeit	Kann mit Konflikten umgehen, erkennt sie, geht sie offen an und kann sie lösen Hat den Mut, anderen entgegenzutreten, sie zu kritisieren oder zu tadeln Lässt sich auf eine mögliche Diskussion ein Fragt nach bei Unstimmigkeiten oder bei auftretenden Problemen Wendet psychologisches Geschick an, um unangenehme zwischenmenschliche Situationen zu bewältigen Sucht und findet einen Konsens Kann mit Kritik umgehen und nimmt Korrekturen und Hinweise auf Reagiert differenziert und sachlich und nimmt nicht alles persönlich
Kommunikationsfähigkeit	Kommuniziert offen, direkt und ehrlich Kann angemessen argumentieren und artikulieren Bleibt bei Meinungsverschiedenheiten ruhig und korrekt Kommuniziert sachlich und in einem der Situation angemessenen Tonfall Erklärt den Sinn und Zweck einer Aufgabe oder einer Übung verständlich Sucht das Gespräch Hört aktiv zu Spricht Klartext mit den Unterstellten
Durchsetzungsfähigkeit	Kann sich – auch gegen Widrigkeiten – durchsetzen Entscheidet auch gegen Widerstände Handelt konsequent und zielorientiert Strebt die Zielerreichung an Steht für seine Meinung ein, bringt eigenen Standpunkt ein Kann sich Gehör verschaffen Erzwingt den Erfolg Duldet unkorrektes Verhalten nicht
Verantwortungsübernahme	Übernimmt die Verantwortung und besitzt Verantwortungsbewusstsein Ist ein Vorbild in seiner Erscheinung und seinen Handlungen Sanktioniert das ungebührende Verhalten der Unterstellten Greift korrigierend ein, falls sich ein Unterstellter falsch verhält Führt und behält den Führungsanspruch Steht für den Befehl ein Stellt sich der Situation, auch wenn sie unangenehm ist Trägt die Konsequenzen
Auftreten als Chef / Selbstsicherheit	Besitzt Selbstvertrauen, ist selbstbewusst und weiss, dass er etwas kann Ist ein Leadertyp Strahlt eine gewisse Autorität und Selbstsicherheit aus Hat ein sicheres Auftreten Ist authentisch Stellt sich dem Konflikt und ist kritikfähig Vermittelt Sicherheit Ist ein Vorbild

Anhang 6.5 Eingestufte Wichtigkeit der einzelnen Verhaltensweisen des Basis-Anforderungsprofils (Berufsmilitärs; rangiert)

Dimension	Verhaltensweise	*Rang*	*M*	*SD*
auftret	Ist ein Vorbild	1	3.64	.52
verantw	Übernimmt die Verantwortung und besitzt Verantwortungsbewusstsein	2	3.42	.59
gewisse	Ist zuverlässig und besitzt Pflichtbewusstsein	3	3.37	.67
gewisse	Ist loyal dem Chef und seinen Unterstellten gegenüber	4	3.35	.66
gewisse	Ist verantwortungs- und pflichtbewusst	5	3.33	.68
fuersor	Behandelt seine Unterstellten und die Mitmenschen respektvoll	6	3.28	.69
verantw	Ist ein Vorbild in seiner Erscheinung und seinen Handlungen	7	3.27	.76
kommuni	Kommuniziert offen, direkt und ehrlich	8	3.19	.66
verantw	Steht für den Befehl ein	9	3.19	.78
planung	Denkt, plant und handelt vorausschauend	10	3.15	.71
verantw	Trägt die Konsequenzen	11	3.13	.70
selbstr	Lernt aus seinen Fehlern und zieht den Mehrwert daraus	12	3.10	.63
planung	Setzt Prioritäten (Wichtigkeit und Dringlichkeit)	13	3.08	.74
durchse	Handelt konsequent und zielorientiert	14	3.08	.65
durchse	Strebt die Zielerreichung an	15	3.08	.65
belastb	Ist psychisch belastbar	16	3.08	.56
fuersor	Schaut zuerst für seine Leute, stellt sich & seine Bedürfnisse in den Hintergrund	17	3.08	.79
gewisse	Ist sorgfältig, diszipliniert, zeitgerecht und auftragstreu	18	3.07	.63
leistun	Will eine gute Leistung erzielen	19	3.07	.61
fuersor	Kennt seine Unterstellten	20	3.05	.77
durchse	Duldet unkorrektes Verhalten nicht	21	3.03	.74
auftret	Ist authentisch	22	3.02	.75
selbstr	Ist ehrlich sich selbst und anderen gegenüber	23	3.02	.68
planung	Kann die Konsequenzen seines Handelns abschätzen	24	3.00	.69
fuersor	Ist fürsorglich gegenüber seinen Unterstellten	25	3.00	.71
analyse	Macht eine angepasste Lagebeurteilung	26	2.98	.82
verantw	Greift korrigierend ein, falls sich ein Unterstellter falsch verhält	27	2.98	.65
verantw	Führt und behält den Führungsanspruch	28	2.98	.72
auftret	Stellt sich dem Konflikt und ist kritikfähig	29	2.97	.58
gewisse	Befolgt, besitzt und vertritt gewisse Werte und Normen	30	2.97	.76
leistun	Ist selbständig und zeigt Eigeninitiative	31	2.95	.67
analyse	Kann die Konsequenzen einschätzen und abschätzen	32	2.95	.57
konflik	Kann mit Konflikten umgehen, erkennt sie, geht sie offen an und kann sie lösen	33	2.93	.61
konflik	Kann mit Kritik umgehen und nimmt Korrekturen und Hinweise auf	34	2.92	.65
leistun	Ist selbst von der Sache überzeugt und begeistert und motiviert so seine Leute	35	2.92	.72
analyse	Sucht und findet eine Lösung	36	2.90	.73
belastb	Bleibt auch unter Druck sachlich, ausgeglichen und zuverlässig	37	2.88	.69
verantw	Stellt sich der Situation, auch wenn sie unangenehm ist	38	2.88	.61
offenhe	Denkt positiv und hat eine positive Grundeinstellung	39	2.85	.78
gewisse	Ist integer und hält sich an Regeln	40	2.83	.78
leistun	Ist motiviert, zeigt Einsatz und packt selbst mit an	41	2.83	.69
kommuni	Spricht Klartext mit den Unterstellten	42	2.82	.62
auftret	Strahlt eine gewisse Autorität und Selbstsicherheit aus	43	2.80	.71
durchse	Kann sich – auch gegen Widrigkeiten – durchsetzen	44	2.80	.66
offenhe	Akzeptiert andere Lösungsvorschläge und Meinungen	45	2.78	.67
leistun	Gibt bei Rückschlägen nicht auf	46	2.77	.56
planung	Setzt Prioritäten	47	2.77	.70
selbstr	Weiss, was sein Verhalten bewirkt	48	2.77	.81
leistun	Ist beharrlich und zeigt Durchhaltewillen	49	2.77	.62
verantw	Sanktioniert das ungebührende Verhalten der Unterstellten	50	2.75	.76
auftret	Hat ein sicheres Auftreten	51	2.73	.61
leistun	Hat Ziele, die er erreichen will und behält die Zielerreichung im Auge	52	2.73	.76
konflik	Reagiert differenziert und sachlich und nimmt nicht alles persönlich	53	2.72	.72
durchse	Steht für seine Meinung ein, bringt eigenen Standpunkt ein	54	2.72	.61
planung	Hat und behält den Überblick	55	2.70	.59
auftret	Vermittelt Sicherheit	56	2.70	.74

Anmerkung: N = 60. Scoring der Wichtigkeitseinstufung:
1 = nicht so wichtig; 2 = wichtig; 3 = sehr wichtig; 4 = unabdingbar.

Anhang 6.5 Eingestufte Wichtigkeit der einzelnen Verhaltensweisen des Basis-Anforderungsprofils (Berufsmilitärs; rangiert, Fortsetzung)

Dimension	Verhaltensweise	*Rang*	*M*	*SD*
planung	Delegiert und erteilt Aufträge	57	2.70	.77
konflik	Hat den Mut, anderen entgegenzutreten, sie zu kritisieren oder zu tadeln	58	2.69	.62
kommuni	Erklärt den Sinn und Zweck einer Aufgabe oder einer Übung verständlich	59	2.68	.68
auftret	Ist ein Leadertyp	60	2.68	.75
fuersor	Nimmt sich Zeit für die Unterstellten	61	2.68	.62
durchse	Entscheidet auch gegen Widerstände	62	2.67	.73
selbstr	Hinterfragt sein Handeln	63	2.67	.73
konflik	Fragt nach bei Unstimmigkeiten oder bei auftretenden Problemen	64	2.67	.71
planung	Nimmt sich genügend Zeit für die Planung & Vorbereitung der Ausbildungsinhalte	65	2.67	.77
analyse	Kann vernetzt denken und erkennt Zusammenhänge	66	2.65	.73
analyse	Kann aus einer Fülle von Informationen die wichtigen erkennen	67	2.65	.66
teamfae	Ist kontaktfreudig und geht auf Menschen zu	68	2.63	.78
kommuni	Kommuniziert sachlich und in einem der Situation angemessenen Tonfall	69	2.61	.72
offenhe	Reagiert flexibel und anpassungsfähig	70	2.60	.64
kommuni	Hört aktiv zu	71	2.60	.62
fuersor	Erkennt die Probleme seiner Leute, geht darauf ein und kann Lösungen anbieten	72	2.60	.69
belastb	Handelt ruhig und überlegt	73	2.60	.62
teamfae	Kommuniziert mit seinen Kameraden um ans Ziel zu kommen	74	2.60	.83
teamfae	Kann sich ein- und unterordnen	75	2.60	.72
offenhe	Passt seinen Führungsstil bewusst an	76	2.58	.65
auftret	Besitzt Selbstvertrauen, ist selbstbewusst und weiss, dass er etwas kann	77	2.57	.59
planung	Geht systematisch und strukturiert vor	78	2.57	.70
durchse	Kann sich Gehör verschaffen	79	2.55	.62
belastb	Hat seine Emotionen – positive wie negative – unter Kontrolle	80	2.53	.81
belastb	Ist bei Überraschungen nicht überfordert	81	2.48	.70
kommuni	Bleibt bei Meinungsverschiedenheiten ruhig und korrekt	82	2.47	.72
analyse	Besitzt eine schnelle Auffassungsgabe	83	2.47	.54
gewisse	Macht auf kritische Punkte aufmerksam	84	2.45	.65
offenhe	Ist offen für Neues und für Verbesserungsvorschläge	85	2.45	.72
kommuni	Kann angemessen argumentieren und artikulieren	86	2.44	.53
teamfae	Vermittelt weiter, wenn er selbst nicht helfen kann	87	2.43	.74
teamfae	Integriert sich in eine Gruppe	88	2.43	.65
teamfae	Kann mit unterschiedlichsten Personen angemessen umgehen	89	2.43	.62
offenhe	Ist bei der Zielerreichung flexibel	90	2.40	.76
fuersor	Verfügt über Einfühlungsvermögen	91	2.40	.67
selbstr	Bleibt realistisch und ist bescheiden	92	2.38	.67
analyse	Begreift neue Sachen schnell und kann sie umsetzen	93	2.38	.58
teamfae	Stellt Gruppenkohäsion her und integriert alle ins Team	94	2.37	.64
selbstr	Ist selbstkritisch und schätzt sich realistisch ein	95	2.35	.71
offenhe	Kann sich in verschiedene Situationen einfühlen	96	2.35	.63
fuersor	Kann sich in andere Menschen einfühlen	97	2.35	.63
selbstr	Kennt seine Schwächen und kann diese auf eine gute Art kompensieren	98	2.34	.63
belastb	Kann physisch mithalten, ist leistungsfähig, ist körperlich fit	99	2.33	.75
kommuni	Sucht das Gespräch	100	2.33	.71
belastb	Besitzt Frustrationstoleranz	101	2.33	.73
offenhe	Sucht nach Alternativen, ist innovativ, kreativ und hat Ideen	102	2.30	.62
selbstr	Ist sich seiner Stärken bewusst	103	2.27	.61
belastb	Zeigt eine konstante Arbeitsleistung	104	2.27	.71
teamfae	Bietet seine Hilfe an	105	2.25	.68
durchse	Erzwingt den Erfolg	106	2.23	.83
analyse	Verfügt über eine hohe Konzentrationsfähigkeit und ein gutes Gedächtnis	107	2.23	.56
konflik	Wendet bei unangenehmer zwischenmenschl. Situation psycholog. Geschick an	108	2.22	.85
konflik	Sucht und findet einen Konsens	109	2.18	.75
gewisse	Hinterfragt die an ihn gestellten Aufträge und überprüft den erhaltenen Befehl	110	2.15	.78
leistun	Meldet sich freiwillig und ist bereit für Mehrarbeit	111	2.05	.72
konflik	Lässt sich auf eine mögliche Diskussion ein	112	1.91	.73

Anmerkung: N = 60. Scoring der Wichtigkeitseinstufung:
1 = nicht so wichtig; 2 = wichtig; 3 = sehr wichtig; 4 = unabdingbar.

Anhang 6.6 Eingestufte Wichtigkeit der einzelnen Verhaltensweisen des Basis-Anforderungsprofils (Berufsmilitärs; gruppiert, rangiert)

Dimension	Verhaltensweise	*Rang*	*M*	*SD*
analyse	Macht eine angepasste Lagebeurteilung	26	2.98	.82
analyse	Kann die Konsequenzen einschätzen und abschätzen	32	2.95	.57
analyse	Sucht und findet eine Lösung	36	2.90	.73
analyse	Kann vernetzt denken und erkennt Zusammenhänge	66	2.65	.73
analyse	Kann aus einer Fülle von Informationen die wichtigen erkennen	67	2.65	.66
analyse	Besitzt eine schnelle Auffassungsgabe	83	2.47	.54
analyse	Begreift neue Sachen schnell und kann sie umsetzen	93	2.38	.58
analyse	Verfügt über eine hohe Konzentrationsfähigkeit und ein gutes Gedächtnis	107	2.23	.56
auftret	Ist ein Vorbild	1	3.64	.52
auftret	Ist authentisch	22	3.02	.75
auftret	Stellt sich dem Konflikt und ist kritikfähig	29	2.97	.58
auftret	Strahlt eine gewisse Autorität und Selbstsicherheit aus	43	2.80	.71
auftret	Hat ein sicheres Auftreten	51	2.73	.61
auftret	Vermittelt Sicherheit	56	2.70	.74
auftret	Ist ein Leadertyp	60	2.68	.75
auftret	Besitzt Selbstvertrauen, ist selbstbewusst und weiss, dass er etwas kann	77	2.57	.59
belastb	Ist psychisch belastbar	16	3.08	.56
belastb	Bleibt auch unter Druck sachlich, ausgeglichen und zuverlässig	37	2.88	.69
belastb	Handelt ruhig und überlegt	73	2.60	.62
belastb	Hat seine Emotionen – positive wie negative – unter Kontrolle	80	2.53	.81
belastb	Ist bei Überraschungen nicht überfordert	81	2.48	.70
belastb	Kann physisch mithalten, ist leistungsfähig, ist körperlich fit	99	2.33	.75
belastb	Besitzt Frustrationstoleranz	101	2.33	.73
belastb	Zeigt eine konstante Arbeitsleistung	104	2.27	.71
durchse	Handelt konsequent und zielorientiert	14	3.08	.65
durchse	Strebt die Zielerreichung an	15	3.08	.65
durchse	Duldet unkorrektes Verhalten nicht	21	3.03	.74
durchse	Kann sich – auch gegen Widrigkeiten – durchsetzen	44	2.80	.66
durchse	Steht für seine Meinung ein, bringt eigenen Standpunkt ein	54	2.72	.61
durchse	Entscheidet auch gegen Widerstände	62	2.67	.73
durchse	Kann sich Gehör verschaffen	79	2.55	.62
durchse	Erzwingt den Erfolg	106	2.23	.83
fuersor	Behandelt seine Unterstellten und die Mitmenschen respektvoll	6	3.28	.69
fuersor	Schaut zuerst für seine Leute, stellt sich & seine Bedürfnisse in den Hintergrund	17	3.08	.79
fuersor	Kennt seine Unterstellten	20	3.05	.77
fuersor	Ist fürsorglich gegenüber seinen Unterstellten	25	3.00	.71
fuersor	Nimmt sich Zeit für die Unterstellten	61	2.68	.62
fuersor	Erkennt die Probleme seiner Leute, geht darauf ein und kann Lösungen anbieten	72	2.60	.69
fuersor	Verfügt über Einfühlungsvermögen	91	2.40	.67
fuersor	Kann sich in andere Menschen einfühlen	97	2.35	.63
gewisse	Ist zuverlässig und besitzt Pflichtbewusstsein	3	3.37	.67
gewisse	Ist loyal dem Chef und seinen Unterstellten gegenüber	4	3.35	.66
gewisse	Ist verantwortungs- und pflichtbewusst	5	3.33	.68
gewisse	Ist sorgfältig, diszipliniert, zeitgerecht und auftragstreu	18	3.07	.63
gewisse	Befolgt, besitzt und vertritt gewisse Werte und Normen	30	2.97	.76
gewisse	Ist integer und hält sich an Regeln	40	2.83	.78
gewisse	Macht auf kritische Punkte aufmerksam	84	2.45	.65
gewisse	Hinterfragt die an ihn gestellten Aufträge und überprüft den erhaltenen Befehl	110	2.15	.78
kommuni	Kommuniziert offen, direkt und ehrlich	8	3.19	.66
kommuni	Spricht Klartext mit den Unterstellten	42	2.82	.62
kommuni	Erklärt den Sinn und Zweck einer Aufgabe oder einer Übung verständlich	59	2.68	.68
kommuni	Kommuniziert sachlich und in einem der Situation angemessenen Tonfall	69	2.61	.72
kommuni	Hört aktiv zu	71	2.60	.62
kommuni	Bleibt bei Meinungsverschiedenheiten ruhig und korrekt	82	2.47	.72
kommuni	Kann angemessen argumentieren und artikulieren	86	2.44	.53
kommuni	Sucht das Gespräch	100	2.33	.71

Anmerkung: N = 60. Scoring der Wichtigkeitseinstufung:
1 = nicht so wichtig; 2 = wichtig; 3 = sehr wichtig; 4 = unabdingbar.

Anhang 6.6 Eingestufte Wichtigkeit der einzelnen Verhaltensweisen des Basis-Anforderungsprofils (Berufsmilitärs; gruppiert, rangiert, Fortsetzung)

Dimension	Verhaltensweise	*Rang*	*M*	*SD*
konflik	Kann mit Konflikten umgehen, erkennt sie, geht sie offen an und kann sie lösen	33	2.93	.61
konflik	Kann mit Kritik umgehen und nimmt Korrekturen und Hinweise auf	34	2.92	.65
konflik	Reagiert differenziert und sachlich und nimmt nicht alles persönlich	53	2.72	.72
konflik	Hat den Mut, anderen entgegenzutreten, sie zu kritisieren oder zu tadeln	58	2.69	.62
konflik	Fragt nach bei Unstimmigkeiten oder bei auftretenden Problemen	64	2.67	.71
konflik	Wendet bei unangenehmer zwischenmenschl. Situation psycholog. Geschick an	108	2.22	.85
konflik	Sucht und findet einen Konsens	109	2.18	.75
konflik	Lässt sich auf eine mögliche Diskussion ein	112	1.91	.73
leistun	Will eine gute Leistung erzielen	19	3.07	.61
leistun	Ist selbständig und zeigt Eigeninitiative	31	2.95	.67
leistun	Ist selbst von der Sache überzeugt und begeistert und motiviert so seine Leute	35	2.92	.72
leistun	Ist motiviert, zeigt Einsatz und packt selbst mit an	41	2.83	.69
leistun	Gibt bei Rückschlägen nicht auf	46	2.77	.56
leistun	Ist beharrlich und zeigt Durchhaltewillen	49	2.77	.62
leistun	Hat Ziele, die er erreichen will und behält die Zielerreichung im Auge	52	2.73	.76
leistun	Meldet sich freiwillig und ist bereit für Mehrarbeit	111	2.05	.72
offenhe	Denkt positiv und hat eine positive Grundeinstellung	39	2.85	.78
offenhe	Akzeptiert andere Lösungsvorschläge und Meinungen	45	2.78	.67
offenhe	Reagiert flexibel und anpassungsfähig	70	2.60	.64
offenhe	Passt seinen Führungsstil bewusst an	76	2.58	.65
offenhe	Ist offen für Neues und für Verbesserungsvorschläge	85	2.45	.72
offenhe	Ist bei der Zielerreichung flexibel	90	2.40	.76
offenhe	Kann sich in verschiedene Situationen einfühlen	96	2.35	.63
offenhe	Sucht nach Alternativen, ist innovativ, kreativ und hat Ideen	102	2.30	.62
planung	Denkt, plant und handelt vorausschauend	10	3.15	.71
planung	Setzt Prioritäten (Wichtigkeit und Dringlichkeit)	13	3.08	.74
planung	Kann die Konsequenzen seines Handelns abschätzen	24	3.00	.69
planung	Setzt Prioritäten	47	2.77	.70
planung	Hat und behält den Überblick	55	2.70	.59
planung	Delegiert und erteilt Aufträge	57	2.70	.77
planung	Nimmt sich genügend Zeit für die Planung & Vorbereitung der Ausbildungsinhalte	65	2.67	.77
planung	Geht systematisch und strukturiert vor	78	2.57	.70
selbstr	Lernt aus seinen Fehlern und zieht den Mehrwert daraus	12	3.10	.63
selbstr	Ist ehrlich sich selbst und anderen gegenüber	23	3.02	.68
selbstr	Weiss, was sein Verhalten bewirkt	48	2.77	.81
selbstr	Hinterfragt sein Handeln	63	2.67	.73
selbstr	Bleibt realistisch und ist bescheiden	92	2.38	.67
selbstr	Ist selbstkritisch und schätzt sich realistisch ein	95	2.35	.71
selbstr	Kennt seine Schwächen und kann diese auf eine gute Art kompensieren	98	2.34	.63
selbstr	Ist sich seiner Stärken bewusst	103	2.27	.61
teamfae	Ist kontaktfreudig und geht auf Menschen zu	68	2.63	.78
teamfae	Kommuniziert mit seinen Kameraden um ans Ziel zu kommen	74	2.60	.83
teamfae	Kann sich ein- und unterordnen	75	2.60	.72
teamfae	Vermittelt weiter, wenn er selbst nicht helfen kann	87	2.43	.74
teamfae	Integriert sich in eine Gruppe	88	2.43	.65
teamfae	Kann mit unterschiedlichsten Personen angemessen umgehen	89	2.43	.62
teamfae	Stellt Gruppenkohäsion her und integriert alle ins Team	94	2.37	.64
teamfae	Bietet seine Hilfe an	105	2.25	.68
verantw	Übernimmt die Verantwortung und besitzt Verantwortungsbewusstsein	2	3.42	.59
verantw	Ist ein Vorbild in seiner Erscheinung und seinen Handlungen	7	3.27	.76
verantw	Steht für den Befehl ein	9	3.19	.78
verantw	Trägt die Konsequenzen	11	3.13	.70
verantw	Greift korrigierend ein, falls sich ein Unterstellter falsch verhält	27	2.98	.65
verantw	Führt und behält den Führungsanspruch	28	2.98	.72
verantw	Stellt sich der Situation, auch wenn sie unangenehm ist	38	2.88	.61
verantw	Sanktioniert das ungebührende Verhalten der Unterstellten	50	2.75	.76

Anmerkung: N = 60. Scoring der Wichtigkeitseinstufung:
1 = nicht so wichtig; 2 = wichtig; 3 = sehr wichtig; 4 = unabdingbar.

Anhang 6.7 Eingestufte Wichtigkeit der einzelnen Verhaltensweisen des Basis-Anforderungsprofils (Gruppenführer; rangiert)

Dimension	Verhaltensweise	*Rang*	*M*	*SD*
auftret	Ist ein Vorbild	1	3.44	.81
gewisse	Ist zuverlässig und besitzt Pflichtbewusstsein	2	3.20	.78
verantw	Ist ein Vorbild in seiner Erscheinung und seinen Handlungen	3	3.15	.78
auftret	Strahlt eine gewisse Autorität und Selbstsicherheit aus	4	3.13	.77
verantw	Übernimmt die Verantwortung und besitzt Verantwortungsbewusstsein	5	3.05	.73
verantw	Führt und behält den Führungsanspruch	6	3.05	.78
fuersor	Behandelt seine Unterstellten und die Mitmenschen respektvoll	7	3.02	.83
gewisse	Ist verantwortungs- und pflichtbewusst	8	3.02	.73
belastb	Ist psychisch belastbar	9	3.00	.80
durchse	Kann sich – auch gegen Widrigkeiten – durchsetzen	10	2.98	.73
durchse	Duldet unkorrektes Verhalten nicht	11	2.98	.85
auftret	Hat ein sicheres Auftreten	12	2.96	.72
fuersor	Schaut zuerst für seine Leute, stellt sich & seine Bedürfnisse in den Hintergrund	13	2.96	.86
leistun	Ist beharrlich und zeigt Durchhaltewillen	14	2.95	.76
leistun	Gibt bei Rückschlägen nicht auf	15	2.93	.81
kommuni	Spricht Klartext mit den Unterstellten	16	2.93	.96
teamfae	Kommuniziert mit seinen Kameraden um ans Ziel zu kommen	17	2.89	.85
organis	Denkt, plant und handelt vorausschauend	18	2.87	.72
gewisse	Ist loyal dem Chef und seiner Gruppe gegenüber	19	2.87	.92
auftret	Besitzt Selbstvertrauen, ist selbstbewusst und weiss, dass er etwas kann	20	2.85	.78
organis	Hat und behält den Überblick	21	2.85	.68
verantw	Trägt die Konsequenzen	22	2.85	.87
durchse	Strebt die Zielerreichung an	23	2.82	.80
analyse	Kann aus einer Fülle von Informationen die wichtigen erkennen	24	2.82	.55
gewisse	Ist sorgfältig, diszipliniert, zeitgerecht und auftragstreu	25	2.80	.78
verantw	Greift korrigierend ein, falls sich ein Unterstellter falsch verhält	26	2.80	.70
konflik	Reagiert differenziert und sachlich und nimmt nicht alles persönlich	27	2.80	.76
konflik	Kann mit Konflikten umgehen, erkennt sie, geht sie offen an und kann sie lösen	28	2.78	.71
kommuni	Kommuniziert offen, direkt und ehrlich	29	2.78	.88
teamfae	Kann mit unterschiedlichsten Personen angemessen umgehen	30	2.78	.81
belastb	Bleibt auch unter Druck sachlich, ausgeglichen und zuverlässig	31	2.78	.66
organis	Setzt Prioritäten (Wichtigkeit und Dringlichkeit)	32	2.78	.74
belastb	Kann physisch mithalten, ist leistungsfähig, ist körperlich fit	33	2.76	.94
leistun	Will eine gute Leistung erzielen	34	2.76	.88
konflik	Hat den Mut, anderen entgegenzutreten, sie zu kritisieren oder zu tadeln	35	2.76	.77
leistun	Ist selbständig und zeigt Eigeninitiative	36	2.75	.89
organis	Setzt Prioritäten	37	2.73	.62
fuersor	Ist fürsorglich gegenüber seiner Gruppe	38	2.71	.83
durchse	Entscheidet auch gegen Widerstände	39	2.71	.74
organis	Delegiert und erteilt Aufträge	40	2.71	.69
selbstr	Lernt aus seinen Fehlern und zieht den Mehrwert daraus	41	2.71	.90
selbstr	Weiss, was sein Verhalten bewirkt	42	2.71	.71
belastb	Handelt ruhig und überlegt	43	2.69	.74
kommuni	Erklärt den Sinn und Zweck einer Aufgabe oder einer Übung verständlich	44	2.69	.86
belastb	Ist bei Überraschungen nicht überfordert	45	2.69	.77
auftret	Vermittelt Sicherheit	46	2.69	.84
fuersor	Erkennt die Probleme seiner Leute, geht darauf ein und kann Lösungen anbieten	47	2.69	.74
leistun	Ist motiviert, zeigt Einsatz und packt selbst mit an	48	2.67	.82
auftret	Ist ein Leadertyp	49	2.67	.94
durchse	Handelt konsequent und zielorientiert	50	2.67	.70
offenhe	Reagiert flexibel und anpassungsfähig	51	2.67	.72
fuersor	Kennt seine Unterstellten	52	2.67	.94
analyse	Kann vernetzt denken und erkennt Zusammenhänge	53	2.67	.70
gewisse	Befolgt, besitzt und vertritt gewisse Werte und Normen	54	2.65	.78
teamfae	Stellt Gruppenkohäsion her und integriert alle ins Team	55	2.65	.78
leistun	Ist selbst von der Sache überzeugt und begeistert und motiviert so seine Leute	56	2.65	.87

Anmerkung: N = 55. Scoring der Wichtigkeitseinstufung:
1 = nicht so wichtig; 2 = wichtig; 3 = sehr wichtig; 4 = unabdingbar.

Anhang 6.7 Eingestufte Wichtigkeit der einzelnen Verhaltensweisen des Basis-Anforderungsprofils (Gruppenführer; rangiert, Fortsetzung)

Dimension	Verhaltensweise	*Rang*	*M*	*SD*
selbstr	Ist sich seiner Stärken bewusst	57	2.65	.67
durchse	Steht für seine Meinung ein, bringt eigenen Standpunkt ein	58	2.65	.70
auftret	Stellt sich dem Konflikt und ist kritikfähig	59	2.65	.75
kommuni	Kann angemessen argumentieren und artikulieren	60	2.64	.73
kommuni	Bleibt bei Meinungsverschiedenheiten ruhig und korrekt	61	2.64	.75
organis	Kann die Konsequenzen seines Handelns abschätzen	62	2.63	.62
leistun	Hat Ziele, die er erreichen will und behält die Zielerreichung im Auge	63	2.60	.87
teamfae	Bietet seine Hilfe an	64	2.60	.87
teamfae	Vermittelt weiter, wenn er selbst nicht helfen kann	65	2.60	.63
analyse	Kann die Konsequenzen einschätzen und abschätzen	66	2.58	.66
verantw	Steht für den Befehl ein	67	2.58	.88
teamfae	Kann sich ein- und unterordnen	68	2.58	.81
belastb	Hat seine Emotionen – positive wie negative – unter Kontrolle	69	2.56	.88
selbstr	Ist ehrlich sich selbst und anderen gegenüber	70	2.55	.72
verantw	Sanktioniert das ungebührende Verhalten der Unterstellten	71	2.55	.74
teamfae	Integriert sich in eine Gruppe	72	2.53	.72
organis	Nimmt sich genügend Zeit für die Planung & Vorbereitung der Ausbildungsinhalte	73	2.53	.74
belastb	Zeigt eine konstante Arbeitsleistung	74	2.53	.66
analyse	Begreift neue Sachen schnell und kann sie umsetzen	75	2.53	.66
organis	Geht systematisch und strukturiert vor	76	2.53	.74
durchse	Kann sich Gehör verschaffen	77	2.53	.86
fuersor	Nimmt sich Zeit für die Unterstellten	78	2.51	.74
konflik	Kann mit Kritik umgehen und nimmt Korrekturen und Hinweise auf	79	2.51	.66
teamfae	Ist kontaktfreudig und geht auf Menschen zu	80	2.49	.84
belastb	Besitzt Frustrationstoleranz	81	2.48	.84
offenhe	Akzeptiert andere Lösungsvorschläge und Meinungen	82	2.47	.84
offenhe	Ist bei der Zielerreichung flexibel	83	2.47	.79
offenhe	Denkt positiv und hat eine positive Grundeinstellung	84	2.45	.96
analyse	Verfügt über eine hohe Konzentrationsfähigkeit und ein gutes Gedächtnis	85	2.45	.79
kommuni	Hört aktiv zu	86	2.45	.63
verantw	Stellt sich der Situation, auch wenn sie unangenehm ist	87	2.45	.81
analyse	Sucht und findet eine Lösung	88	2.44	.57
konflik	Fragt nach bei Unstimmigkeiten oder bei auftretenden Problemen	89	2.44	.74
kommuni	Kommuniziert sachlich und in einem der Situation angemessenen Tonfall	90	2.44	.71
offenhe	Passt seinen Führungsstil bewusst an	91	2.44	.71
analyse	Macht eine angepasste Lagebeurteilung	92	2.38	.78
analyse	Besitzt eine schnelle Auffassungsgabe	93	2.38	.73
konflik	Sucht und findet einen Konsens	94	2.37	.71
offenhe	Ist offen für Neues und für Verbesserungsvorschläge	95	2.36	.70
leistun	Meldet sich freiwillig und ist bereit für Mehrarbeit	96	2.36	.91
gewisse	Hinterfragt die an ihn gestellten Aufträge und überprüft den erhaltenen Befehl	97	2.36	.73
gewisse	Macht auf kritische Punkte aufmerksam	98	2.35	.70
konflik	Wendet bei unangenehmer zwischenmenschl. Situation psycholog. Geschick an	99	2.35	.78
gewisse	Ist integer und hält sich an Regeln	100	2.33	.79
selbstr	Bleibt realistisch und ist bescheiden	101	2.33	.82
selbstr	Hinterfragt sein Handeln	102	2.29	.81
offenhe	Sucht nach Alternativen, ist innovativ, kreativ und hat Ideen	103	2.27	.78
fuersor	Kann sich in andere Menschen einfühlen	104	2.27	.91
selbstr	Kennt seine Schwächen und kann diese auf eine gute Art kompensieren	105	2.24	.69
offenhe	Kann sich in verschiedene Situationen einfühlen	106	2.22	.63
selbstr	Ist selbstkritisch und schätzt sich realistisch ein	107	2.20	.70
fuersor	Verfügt über Einfühlungsvermögen	108	2.16	.86
auftret	Ist authentisch	109	2.15	.78
kommuni	Sucht das Gespräch	110	2.07	.74
durchse	Erzwingt den Erfolg	111	2.05	.83
konflik	Lässt sich auf eine mögliche Diskussion ein	112	1.87	.72

Anmerkung: N = 55. Scoring der Wichtigkeitseinstufung:
1 = nicht so wichtig; 2 = wichtig; 3 = sehr wichtig; 4 = unabdingbar.

Anhang 6.8 Eingestufte Wichtigkeit der einzelnen Verhaltensweisen des Basis-Anforderungsprofils (Gruppenführer; gruppiert, rangiert)

Dimension	Verhaltensweise	*Rang*	*M*	*SD*
analyse	Kann aus einer Fülle von Informationen die wichtigen erkennen	24	2.82	.55
analyse	Kann vernetzt denken und erkennt Zusammenhänge	53	2.67	.70
analyse	Kann die Konsequenzen einschätzen und abschätzen	66	2.58	.66
analyse	Begreift neue Sachen schnell und kann sie umsetzen	75	2.53	.66
analyse	Verfügt über eine hohe Konzentrationsfähigkeit und ein gutes Gedächtnis	85	2.45	.79
analyse	Sucht und findet eine Lösung	88	2.44	.57
analyse	Macht eine angepasste Lagebeurteilung	92	2.38	.78
analyse	Besitzt eine schnelle Auffassungsgabe	93	2.38	.73
auftret	Ist ein Vorbild	1	3.44	.81
auftret	Strahlt eine gewisse Autorität und Selbstsicherheit aus	4	3.13	.77
auftret	Hat ein sicheres Auftreten	12	2.96	.72
auftret	Besitzt Selbstvertrauen, ist selbstbewusst und weiss, dass er etwas kann	20	2.85	.78
auftret	Vermittelt Sicherheit	46	2.69	.84
auftret	Ist ein Leadertyp	49	2.67	.94
auftret	Stellt sich dem Konflikt und ist kritikfähig	59	2.65	.75
auftret	Ist authentisch	109	2.15	.78
belastb	Ist psychisch belastbar	9	3.00	.80
belastb	Bleibt auch unter Druck sachlich, ausgeglichen und zuverlässig	31	2.78	.66
belastb	Kann physisch mithalten, ist leistungsfähig, ist körperlich fit	33	2.76	.94
belastb	Handelt ruhig und überlegt	43	2.69	.74
belastb	Ist bei Überraschungen nicht überfordert	45	2.69	.77
belastb	Hat seine Emotionen – positive wie negative – unter Kontrolle	69	2.56	.88
belastb	Zeigt eine konstante Arbeitsleistung	74	2.53	.66
belastb	Besitzt Frustrationstoleranz	81	2.48	.84
durchse	Kann sich – auch gegen Widrigkeiten – durchsetzen	10	2.98	.73
durchse	Duldet unkorrektes Verhalten nicht	11	2.98	.85
durchse	Strebt die Zielerreichung an	23	2.82	.80
durchse	Entscheidet auch gegen Widerstände	39	2.71	.74
durchse	Handelt konsequent und zielorientiert	50	2.67	.70
durchse	Steht für seine Meinung ein, bringt eigenen Standpunkt ein	58	2.65	.70
durchse	Kann sich Gehör verschaffen	77	2.53	.86
durchse	Erzwingt den Erfolg	111	2.05	.83
fuersor	Behandelt seine Unterstellten und die Mitmenschen respektvoll	7	3.02	.83
fuersor	Schaut zuerst für seine Leute, stellt sich & seine Bedürfnisse in den Hintergrund	13	2.96	.86
fuersor	Ist fürsorglich gegenüber seiner Gruppe	38	2.71	.83
fuersor	Erkennt die Probleme seiner Leute, geht darauf ein und kann Lösungen anbieten	47	2.69	.74
fuersor	Kennt seine Unterstellten	52	2.67	.94
fuersor	Nimmt sich Zeit für die Unterstellten	78	2.51	.74
fuersor	Kann sich in andere Menschen einfühlen	104	2.27	.91
fuersor	Verfügt über Einfühlungsvermögen	108	2.16	.86
gewisse	Ist zuverlässig und besitzt Pflichtbewusstsein	2	3.20	.78
gewisse	Ist verantwortungs- und pflichtbewusst	8	3.02	.73
gewisse	Ist loyal dem Chef und seiner Gruppe gegenüber	19	2.87	.92
gewisse	Ist sorgfältig, diszipliniert, zeitgerecht und auftragstreu	25	2.80	.78
gewisse	Befolgt, besitzt und vertritt gewisse Werte und Normen	54	2.65	.78
gewisse	Hinterfragt die an ihn gestellten Aufträge und überprüft den erhaltenen Befehl	97	2.36	.73
gewisse	Macht auf kritische Punkte aufmerksam	98	2.35	.70
gewisse	Ist integer und hält sich an Regeln	100	2.33	.79
kommuni	Spricht Klartext mit den Unterstellten	16	2.93	.96
kommuni	Kommuniziert offen, direkt und ehrlich	29	2.78	.88
kommuni	Erklärt den Sinn und Zweck einer Aufgabe oder einer Übung verständlich	44	2.69	.86
kommuni	Kann angemessen argumentieren und artikulieren	60	2.64	.73
kommuni	Bleibt bei Meinungsverschiedenheiten ruhig und korrekt	61	2.64	.75
kommuni	Hört aktiv zu	86	2.45	.63
kommuni	Kommuniziert sachlich und in einem der Situation angemessenen Tonfall	90	2.44	.71
kommuni	Sucht das Gespräch	110	2.07	.74

Anmerkung: N = 55. Scoring der Wichtigkeitseinstufung:
1 = nicht so wichtig; 2 = wichtig; 3 = sehr wichtig; 4 = unabdingbar.

Anhang 6.8 Eingestufte Wichtigkeit der einzelnen Verhaltensweisen des Basis-Anforderungsprofils (Gruppenführer; gruppiert, rangiert, Fortsetzung)

Dimension	Verhaltensweise	*Rang*	*M*	*SD*
konflik	Reagiert differenziert und sachlich und nimmt nicht alles persönlich	27	2.80	.76
konflik	Kann mit Konflikten umgehen, erkennt sie, geht sie offen an und kann sie lösen	28	2.78	.71
konflik	Hat den Mut, anderen entgegenzutreten, sie zu kritisieren oder zu tadeln	35	2.76	.77
konflik	Kann mit Kritik umgehen und nimmt Korrekturen und Hinweise auf	79	2.51	.66
konflik	Fragt nach bei Unstimmigkeiten oder bei auftretenden Problemen	89	2.44	.74
konflik	Sucht und findet einen Konsens	94	2.37	.71
konflik	Wendet bei unangenehmer zwischenmenschl. Situation psycholog. Geschick an	99	2.35	.78
konflik	Lässt sich auf eine mögliche Diskussion ein	112	1.87	.72
leistun	Ist beharrlich und zeigt Durchhaltewillen	14	2.95	.76
leistun	Gibt bei Rückschlägen nicht auf	15	2.93	.81
leistun	Will eine gute Leistung erzielen	34	2.76	.88
leistun	Ist selbständig und zeigt Eigeninitiative	36	2.75	.89
leistun	Ist motiviert, zeigt Einsatz und packt selbst mit an	48	2.67	.82
leistun	Ist selbst von der Sache überzeugt und begeistert und motiviert so seine Leute	56	2.65	.87
leistun	Hat Ziele, die er erreichen will und behält die Zielerreichung im Auge	63	2.60	.87
leistun	Meldet sich freiwillig und ist bereit für Mehrarbeit	96	2.36	.91
offenhe	Reagiert flexibel und anpassungsfähig	51	2.67	.72
offenhe	Akzeptiert andere Lösungsvorschläge und Meinungen	82	2.47	.84
offenhe	Ist bei der Zielerreichung flexibel	83	2.47	.79
offenhe	Denkt positiv und hat eine positive Grundeinstellung	84	2.45	.96
offenhe	Passt seinen Führungsstil bewusst an	91	2.44	.71
offenhe	Ist offen für Neues und für Verbesserungsvorschläge	95	2.36	.70
offenhe	Sucht nach Alternativen, ist innovativ, kreativ und hat Ideen	103	2.27	.78
offenhe	Kann sich in verschiedene Situationen einfühlen	106	2.22	.63
organis	Denkt, plant und handelt vorausschauend	18	2.87	.72
organis	Hat und behält den Überblick	21	2.85	.68
organis	Setzt Prioritäten (Wichtigkeit und Dringlichkeit)	32	2.78	.74
organis	Setzt Prioritäten	37	2.73	.62
organis	Delegiert und erteilt Aufträge	40	2.71	.69
organis	Kann die Konsequenzen seines Handelns abschätzen	62	2.63	.62
organis	Nimmt sich genügend Zeit für die Planung & Vorbereitung der Ausbildungsinhalte	73	2.53	.74
organis	Geht systematisch und strukturiert vor	76	2.53	.74
selbstr	Lernt aus seinen Fehlern und zieht den Mehrwert daraus	41	2.71	.90
selbstr	Weiss, was sein Verhalten bewirkt	42	2.71	.71
selbstr	Ist sich seiner Stärken bewusst	57	2.65	.67
selbstr	Ist ehrlich sich selbst und anderen gegenüber	70	2.55	.72
selbstr	Bleibt realistisch und ist bescheiden	101	2.33	.82
selbstr	Hinterfragt sein Handeln	102	2.29	.81
selbstr	Kennt seine Schwächen und kann diese auf eine gute Art kompensieren	105	2.24	.69
selbstr	Ist selbstkritisch und schätzt sich realistisch ein	107	2.20	.70
teamfae	Kommuniziert mit seinen Kameraden um ans Ziel zu kommen	17	2.89	.85
teamfae	Kann mit unterschiedlichsten Personen angemessen umgehen	30	2.78	.81
teamfae	Stellt Gruppenkohäsion her und integriert alle ins Team	55	2.65	.78
teamfae	Bietet seine Hilfe an	64	2.60	.87
teamfae	Vermittelt weiter, wenn er selbst nicht helfen kann	65	2.60	.63
teamfae	Kann sich ein- und unterordnen	68	2.58	.81
teamfae	Integriert sich in eine Gruppe	72	2.53	.72
teamfae	Ist kontaktfreudig und geht auf Menschen zu	80	2.49	.84
verantw	Ist ein Vorbild in seiner Erscheinung und seinen Handlungen	3	3.15	.78
verantw	Übernimmt die Verantwortung und besitzt Verantwortungsbewusstsein	5	3.05	.73
verantw	Führt und behält den Führungsanspruch	6	3.05	.78
verantw	Trägt die Konsequenzen	22	2.85	.87
verantw	Greift korrigierend ein, falls sich ein Unterstellter falsch verhält	26	2.80	.70
verantw	Steht für den Befehl ein	67	2.58	.88
verantw	Sanktioniert das ungebührende Verhalten der Unterstellten	71	2.55	.74
verantw	Stellt sich der Situation, auch wenn sie unangenehm ist	87	2.45	.81

Anmerkung: N = 55. Scoring der Wichtigkeitseinstufung:
1 = nicht so wichtig; 2 = wichtig; 3 = sehr wichtig; 4 = unabdingbar.

Anhang 6.9 Eingestufte Wichtigkeit der einzelnen Verhaltensweisen des Basis-Anforderungsprofils (Zugführer; rangiert)

Dimension	Verhaltensweise	*Rang*	*M*	*SD*
auftret	Ist ein Vorbild	1	3.59	.54
verantw	Übernimmt die Verantwortung und besitzt Verantwortungsbewusstsein	2	3.41	.64
gewisse	Ist zuverlässig und besitzt Pflichtbewusstsein	3	3.35	.63
verantw	Ist ein Vorbild in seiner Erscheinung und seinen Handlungen	4	3.35	.72
organis	Denkt, plant und handelt vorausschauend	5	3.33	.65
leistun	Will eine gute Leistung erzielen	6	3.29	.67
verantw	Führt und behält den Führungsanspruch	7	3.29	.70
selbstr	Lernt aus seinen Fehlern und zieht den Mehrwert daraus	8	3.27	.63
fuersor	Schaut zuerst für seine Leute, stellt sich & seine Bedürfnisse in den Hintergrund	9	3.27	.72
verantw	Trägt die Konsequenzen	10	3.27	.70
gewisse	Ist verantwortungs- und pflichtbewusst	11	3.25	.63
organis	Setzt Prioritäten (Wichtigkeit und Dringlichkeit)	12	3.25	.63
auftret	Strahlt eine gewisse Autorität und Selbstsicherheit aus	13	3.24	.79
leistun	Ist selbständig und zeigt Eigeninitiative	14	3.22	.70
durchse	Duldet unkorrektes Verhalten nicht	15	3.20	.78
auftret	Hat ein sicheres Auftreten	16	3.16	.67
belastb	Ist psychisch belastbar	17	3.14	.57
verantw	Greift korrigierend ein, falls sich ein Unterstellter falsch verhält	18	3.12	.62
belastb	Bleibt auch unter Druck sachlich, ausgeglichen und zuverlässig	19	3.12	.59
organis	Hat und behält den Überblick	20	3.10	.67
gewisse	Ist loyal dem Chef und seiner Gruppe gegenüber	21	3.10	.73
auftret	Besitzt Selbstvertrauen, ist selbstbewusst und weiss, dass er etwas kann	22	3.08	.74
fuersor	Behandelt seine Unterstellten und die Mitmenschen respektvoll	23	3.08	.77
leistun	Gibt bei Rückschlägen nicht auf	24	3.08	.72
leistun	Ist beharrlich und zeigt Durchhaltewillen	25	3.08	.66
offenhe	Reagiert flexibel und anpassungsfähig	26	3.06	.76
verantw	Steht für den Befehl ein	27	3.06	.70
organis	Kann die Konsequenzen seines Handelns abschätzen	28	3.06	.65
durchse	Kann sich – auch gegen Widrigkeiten – durchsetzen	29	3.04	.77
leistun	Ist motiviert, zeigt Einsatz und packt selbst mit an	30	3.04	.92
organis	Setzt Prioritäten	31	3.04	.77
belastb	Ist bei Überraschungen nicht überfordert	32	3.04	.63
durchse	Entscheidet auch gegen Widerstände	33	3.02	.73
organis	Delegiert und erteilt Aufträge	34	3.02	.76
leistun	Hat Ziele, die er erreichen will und behält die Zielerreichung im Auge	35	3.00	.85
durchse	Strebt die Zielerreichung an	36	3.00	.69
organis	Geht systematisch und strukturiert vor	37	3.00	.82
gewisse	Ist sorgfältig, diszipliniert, zeitgerecht und auftragstreu	38	2.98	.71
auftret	Ist ein Leadertyp	39	2.98	.88
durchse	Handelt konsequent und zielorientiert	40	2.98	.68
durchse	Kann sich Gehör verschaffen	41	2.98	.84
analyse	Kann die Konsequenzen einschätzen und abschätzen	42	2.96	.77
teamfae	Kann mit unterschiedlichsten Personen angemessen umgehen	43	2.96	.69
gewisse	Befolgt, besitzt und vertritt gewisse Werte und Normen	44	2.96	.77
leistun	Ist selbst von der Sache überzeugt und begeistert und motiviert so seine Leute	45	2.96	.80
auftret	Stellt sich dem Konflikt und ist kritikfähig	46	2.96	.69
belastb	Hat seine Emotionen – positive wie negative – unter Kontrolle	47	2.96	.85
selbstr	Weiss, was sein Verhalten bewirkt	48	2.96	.75
teamfae	Kommuniziert mit seinen Kameraden um ans Ziel zu kommen	49	2.96	.72
kommuni	Spricht Klartext mit den Unterstellten	50	2.96	.77
analyse	Macht eine angepasste Lagebeurteilung	51	2.94	.86
analyse	Sucht und findet eine Lösung	52	2.94	.70
kommuni	Kann angemessen argumentieren und artikulieren	53	2.94	.76
kommuni	Erklärt den Sinn und Zweck einer Aufgabe oder einer Übung verständlich	54	2.94	.70
analyse	Kann vernetzt denken und erkennt Zusammenhänge	55	2.94	.73
konflik	Hat den Mut, anderen entgegenzutreten, sie zu kritisieren oder zu tadeln	56	2.92	.80

Anmerkung: N = 51. Scoring der Wichtigkeitseinstufung:
1 = nicht so wichtig; 2 = wichtig; 3 = sehr wichtig; 4 = unabdingbar.

Anhang 6.9 Eingestufte Wichtigkeit der einzelnen Verhaltensweisen des Basis-Anforderungsprofils (Zugführer; rangiert, Fortsetzung)

Dimension	Verhaltensweise	*Rang*	*M*	*SD*
analyse	Kann aus einer Fülle von Informationen die wichtigen erkennen	57	2.92	.77
belastb	Kann physisch mithalten, ist leistungsfähig, ist körperlich fit	58	2.90	.90
kommuni	Kommuniziert offen, direkt und ehrlich	59	2.90	.81
gewisse	Ist integer und hält sich an Regeln	60	2.90	.67
fuersor	Kennt seine Unterstellten	61	2.90	.85
offenhe	Denkt positiv und hat eine positive Grundeinstellung	62	2.90	.78
konflik	Kann mit Kritik umgehen und nimmt Korrekturen und Hinweise auf	63	2.90	.81
verantw	Stellt sich der Situation, auch wenn sie unangenehm ist	64	2.90	.73
auftret	Vermittelt Sicherheit	65	2.88	.71
fuersor	Ist fürsorglich gegenüber seiner Gruppe	66	2.86	.80
konflik	Kann mit Konflikten umgehen, erkennt sie, geht sie offen an und kann sie lösen	67	2.86	.78
fuersor	Erkennt die Probleme seiner Leute, geht darauf ein und kann Lösungen anbieten	68	2.86	.66
durchse	Steht für seine Meinung ein, bringt eigenen Standpunkt ein	69	2.84	.78
kommuni	Bleibt bei Meinungsverschiedenheiten ruhig und korrekt	70	2.82	.71
analyse	Verfügt über eine hohe Konzentrationsfähigkeit und ein gutes Gedächtnis	71	2.82	.77
teamfae	Kann sich ein- und unterordnen	72	2.82	.84
kommuni	Hört aktiv zu	73	2.82	.71
verantw	Sanktioniert das ungebührende Verhalten der Unterstellten	74	2.80	.72
offenhe	Passt seinen Führungsstil bewusst an	75	2.80	.89
selbstr	Ist sich seiner Stärken bewusst	76	2.78	.86
offenhe	Ist bei der Zielerreichung flexibel	77	2.78	.81
belastb	Handelt ruhig und überlegt	78	2.76	.65
offenhe	Akzeptiert andere Lösungsvorschläge und Meinungen	79	2.76	.74
kommuni	Kommuniziert sachlich und in einem der Situation angemessenen Tonfall	80	2.75	.77
selbstr	Ist ehrlich sich selbst und anderen gegenüber	81	2.73	.83
organis	Nimmt sich genügend Zeit für die Planung & Vorbereitung der Ausbildungsinhalte	82	2.73	.72
belastb	Zeigt eine konstante Arbeitsleistung	83	2.73	.92
analyse	Begreift neue Sachen schnell und kann sie umsetzen	84	2.73	.83
teamfae	Vermittelt weiter, wenn er selbst nicht helfen kann	85	2.72	.76
fuersor	Nimmt sich Zeit für die Unterstellten	86	2.71	.83
analyse	Besitzt eine schnelle Auffassungsgabe	87	2.69	.73
durchse	Erzwingt den Erfolg	88	2.69	.88
gewisse	Hinterfragt die an ihn gestellten Aufträge und überprüft den erhaltenen Befehl	89	2.69	.95
konflik	Reagiert differenziert und sachlich und nimmt nicht alles persönlich	90	2.69	.84
auftret	Ist authentisch	91	2.68	.79
teamfae	Ist kontaktfreudig und geht auf Menschen zu	92	2.67	.82
fuersor	Kann sich in andere Menschen einfühlen	93	2.67	.84
teamfae	Stellt Gruppenkohäsion her und integriert alle ins Team	94	2.65	.87
gewisse	Macht auf kritische Punkte aufmerksam	95	2.65	.89
selbstr	Kennt seine Schwächen und kann diese auf eine gute Art kompensieren	96	2.63	.80
konflik	Sucht und findet einen Konsens	97	2.63	.77
belastb	Besitzt Frustrationstoleranz	98	2.63	.77
selbstr	Hinterfragt sein Handeln	99	2.59	.90
konflik	Fragt nach bei Unstimmigkeiten oder bei auftretenden Problemen	100	2.59	.78
fuersor	Verfügt über Einfühlungsvermögen	101	2.59	.80
offenhe	Kann sich in verschiedene Situationen einfühlen	102	2.57	.83
selbstr	Bleibt realistisch und ist bescheiden	103	2.56	.79
offenhe	Ist offen für Neues und für Verbesserungsvorschläge	104	2.53	.76
leistun	Meldet sich freiwillig und ist bereit für Mehrarbeit	105	2.53	1.0
offenhe	Sucht nach Alternativen, ist innovativ, kreativ und hat Ideen	106	2.51	.86
selbstr	Ist selbstkritisch und schätzt sich realistisch ein	107	2.51	.78
konflik	Wendet bei unangenehmer zwischenmenschl. Situation psycholog. Geschick an	108	2.51	.97
teamfae	Bietet seine Hilfe an	109	2.43	.83
kommuni	Sucht das Gespräch	110	2.41	.88
teamfae	Integriert sich in eine Gruppe	111	2.31	.95
konflik	Lässt sich auf eine mögliche Diskussion ein	112	2.24	1.1

Anmerkung: N = 51. Scoring der Wichtigkeitseinstufung:
1 = nicht so wichtig; 2 = wichtig; 3 = sehr wichtig; 4 = unabdingbar.

Anhang 6.10 Eingestufte Wichtigkeit der einzelnen Verhaltensweisen des Basis-Anforderungsprofils (Zugführer; gruppiert, rangiert)

Dimension	Verhaltensweise	*Rang*	*M*	*SD*
analyse	Kann die Konsequenzen einschätzen und abschätzen	42	2.96	.77
analyse	Macht eine angepasste Lagebeurteilung	51	2.94	.86
analyse	Sucht und findet eine Lösung	52	2.94	.70
analyse	Kann vernetzt denken und erkennt Zusammenhänge	55	2.94	.73
analyse	Kann aus einer Fülle von Informationen die wichtigen erkennen	57	2.92	.77
analyse	Verfügt über eine hohe Konzentrationsfähigkeit und ein gutes Gedächtnis	71	2.82	.77
analyse	Begreift neue Sachen schnell und kann sie umsetzen	84	2.73	.83
analyse	Besitzt eine schnelle Auffassungsgabe	87	2.69	.73
auftret	Ist ein Vorbild	1	3.59	.54
auftret	Strahlt eine gewisse Autorität und Selbstsicherheit aus	13	3.24	.79
auftret	Hat ein sicheres Auftreten	16	3.16	.67
auftret	Besitzt Selbstvertrauen, ist selbstbewusst und weiss, dass er etwas kann	22	3.08	.74
auftret	Ist ein Leadertyp	39	2.98	.88
auftret	Stellt sich dem Konflikt und ist kritikfähig	46	2.96	.69
auftret	Vermittelt Sicherheit	65	2.88	.71
auftret	Ist authentisch	91	2.68	.79
belastb	Ist psychisch belastbar	17	3.14	.57
belastb	Bleibt auch unter Druck sachlich, ausgeglichen und zuverlässig	19	3.12	.59
belastb	Ist bei Überraschungen nicht überfordert	32	3.04	.63
belastb	Hat seine Emotionen – positive wie negative – unter Kontrolle	47	2.96	.85
belastb	Kann physisch mithalten, ist leistungsfähig, ist körperlich fit	58	2.90	.90
belastb	Handelt ruhig und überlegt	78	2.76	.65
belastb	Zeigt eine konstante Arbeitsleistung	83	2.73	.92
belastb	Besitzt Frustrationstoleranz	98	2.63	.77
durchse	Duldet unkorrektes Verhalten nicht	15	3.20	.78
durchse	Kann sich – auch gegen Widrigkeiten – durchsetzen	29	3.04	.77
durchse	Entscheidet auch gegen Widerstände	33	3.02	.73
durchse	Strebt die Zielerreichung an	36	3.00	.69
durchse	Handelt konsequent und zielorientiert	40	2.98	.68
durchse	Kann sich Gehör verschaffen	41	2.98	.84
durchse	Steht für seine Meinung ein, bringt eigenen Standpunkt ein	69	2.84	.78
durchse	Erzwingt den Erfolg	88	2.69	.88
fuersor	Schaut zuerst für seine Leute, stellt sich & seine Bedürfnisse in den Hintergrund	9	3.27	.72
fuersor	Behandelt seine Unterstellten und die Mitmenschen respektvoll	23	3.08	.77
fuersor	Kennt seine Unterstellten	61	2.90	.85
fuersor	Ist fürsorglich gegenüber seiner Gruppe	66	2.86	.80
fuersor	Erkennt die Probleme seiner Leute, geht darauf ein und kann Lösungen anbieten	68	2.86	.66
fuersor	Nimmt sich Zeit für die Unterstellten	86	2.71	.83
fuersor	Kann sich in andere Menschen einfühlen	93	2.67	.84
fuersor	Verfügt über Einfühlungsvermögen	101	2.59	.80
gewisse	Ist zuverlässig und besitzt Pflichtbewusstsein	3	3.35	.63
gewisse	Ist verantwortungs- und pflichtbewusst	11	3.25	.63
gewisse	Ist loyal dem Chef und seiner Gruppe gegenüber	21	3.10	.73
gewisse	Ist sorgfältig, diszipliniert, zeitgerecht und auftragstreu	38	2.98	.71
gewisse	Befolgt, besitzt und vertritt gewisse Werte und Normen	44	2.96	.77
gewisse	Ist integer und hält sich an Regeln	60	2.90	.67
gewisse	Hinterfragt die an ihn gestellten Aufträge und überprüft den erhaltenen Befehl	89	2.69	.95
gewisse	Macht auf kritische Punkte aufmerksam	95	2.65	.89
kommuni	Spricht Klartext mit den Unterstellten	50	2.96	.77
kommuni	Kann angemessen argumentieren und artikulieren	53	2.94	.76
kommuni	Erklärt den Sinn und Zweck einer Aufgabe oder einer Übung verständlich	54	2.94	.70
kommuni	Kommuniziert offen, direkt und ehrlich	59	2.90	.81
kommuni	Bleibt bei Meinungsverschiedenheiten ruhig und korrekt	70	2.82	.71
kommuni	Hört aktiv zu	73	2.82	.71
kommuni	Kommuniziert sachlich und in einem der Situation angemessenen Tonfall	80	2.75	.77
kommuni	Sucht das Gespräch	110	2.41	.88

Anmerkung: N = 51. Scoring der Wichtigkeitseinstufung:
1 = nicht so wichtig; 2 = wichtig; 3 = sehr wichtig; 4 = unabdingbar.

Anhang 6.10 Eingestufte Wichtigkeit der einzelnen Verhaltensweisen des Basis-Anforderungsprofils (Zugführer; gruppiert, rangiert, Fortsetzung)

Dimension	Verhaltensweise	*Rang*	*M*	*SD*
konflik	Hat den Mut, anderen entgegenzutreten, sie zu kritisieren oder zu tadeln	56	2.92	.80
konflik	Kann mit Kritik umgehen und nimmt Korrekturen und Hinweise auf	63	2.90	.81
konflik	Kann mit Konflikten umgehen, erkennt sie, geht sie offen an und kann sie lösen	67	2.86	.78
konflik	Reagiert differenziert und sachlich und nimmt nicht alles persönlich	90	2.69	.84
konflik	Sucht und findet einen Konsens	97	2.63	.77
konflik	Fragt nach bei Unstimmigkeiten oder bei auftretenden Problemen	100	2.59	.78
konflik	Wendet bei unangenehmer zwischenmenschl. Situation psycholog. Geschick an	108	2.51	.97
konflik	Lässt sich auf eine mögliche Diskussion ein	112	2.24	1.1
leistun	Will eine gute Leistung erzielen	6	3.29	.67
leistun	Ist selbständig und zeigt Eigeninitiative	14	3.22	.70
leistun	Gibt bei Rückschlägen nicht auf	24	3.08	.72
leistun	Ist beharrlich und zeigt Durchhaltewillen	25	3.08	.66
leistun	Ist motiviert, zeigt Einsatz und packt selbst mit an	30	3.04	.92
leistun	Hat Ziele, die er erreichen will und behält die Zielerreichung im Auge	35	3.00	.85
leistun	Ist selbst von der Sache überzeugt und begeistert und motiviert so seine Leute	45	2.96	.80
leistun	Meldet sich freiwillig und ist bereit für Mehrarbeit	105	2.53	1.0
offenhe	Reagiert flexibel und anpassungsfähig	26	3.06	.76
offenhe	Denkt positiv und hat eine positive Grundeinstellung	62	2.90	.78
offenhe	Passt seinen Führungsstil bewusst an	75	2.80	.89
offenhe	Ist bei der Zielerreichung flexibel	77	2.78	.81
offenhe	Akzeptiert andere Lösungsvorschläge und Meinungen	79	2.76	.74
offenhe	Kann sich in verschiedene Situationen einfühlen	102	2.57	.83
offenhe	Ist offen für Neues und für Verbesserungsvorschläge	104	2.53	.76
offenhe	Sucht nach Alternativen, ist innovativ, kreativ und hat Ideen	106	2.51	.86
organis	Denkt, plant und handelt vorausschauend	5	3.33	.65
organis	Setzt Prioritäten (Wichtigkeit und Dringlichkeit)	12	3.25	.63
organis	Hat und behält den Überblick	20	3.10	.67
organis	Kann die Konsequenzen seines Handelns abschätzen	28	3.06	.65
organis	Setzt Prioritäten	31	3.04	.77
organis	Delegiert und erteilt Aufträge	34	3.02	.76
organis	Geht systematisch und strukturiert vor	37	3.00	.82
organis	Nimmt sich genügend Zeit für die Planung & Vorbereitung der Ausbildungsinhalte	82	2.73	.72
selbstr	Lernt aus seinen Fehlern und zieht den Mehrwert daraus	8	3.27	.63
selbstr	Weiss, was sein Verhalten bewirkt	48	2.96	.75
selbstr	Ist sich seiner Stärken bewusst	76	2.78	.86
selbstr	Ist ehrlich sich selbst und anderen gegenüber	81	2.73	.83
selbstr	Kennt seine Schwächen und kann diese auf eine gute Art kompensieren	96	2.63	.80
selbstr	Hinterfragt sein Handeln	99	2.59	.90
selbstr	Bleibt realistisch und ist bescheiden	103	2.56	.79
selbstr	Ist selbstkritisch und schätzt sich realistisch ein	107	2.51	.78
teamfae	Kann mit unterschiedlichsten Personen angemessen umgehen	43	2.96	.69
teamfae	Kommuniziert mit seinen Kameraden um ans Ziel zu kommen	49	2.96	.72
teamfae	Kann sich ein- und unterordnen	72	2.82	.84
teamfae	Vermittelt weiter, wenn er selbst nicht helfen kann	85	2.72	.76
teamfae	Ist kontaktfreudig und geht auf Menschen zu	92	2.67	.82
teamfae	Stellt Gruppenkohäsion her und integriert alle ins Team	94	2.65	.87
teamfae	Bietet seine Hilfe an	109	2.43	.83
teamfae	Integriert sich in eine Gruppe	111	2.31	.95
verantw	Übernimmt die Verantwortung und besitzt Verantwortungsbewusstsein	2	3.41	.64
verantw	Ist ein Vorbild in seiner Erscheinung und seinen Handlungen	4	3.35	.72
verantw	Führt und behält den Führungsanspruch	7	3.29	.70
verantw	Trägt die Konsequenzen	10	3.27	.70
verantw	Greift korrigierend ein, falls sich ein Unterstellter falsch verhält	18	3.12	.62
verantw	Steht für den Befehl ein	27	3.06	.70
verantw	Stellt sich der Situation, auch wenn sie unangenehm ist	64	2.90	.73
verantw	Sanktioniert das ungebührende Verhalten der Unterstellten	74	2.80	.72

Anmerkung: N = 51. Scoring der Wichtigkeitseinstufung:
1 = nicht so wichtig; 2 = wichtig; 3 = sehr wichtig; 4 = unabdingbar.

Anhang 6.11 Prototypenrating der Acts zur Dimension Durchsetzungsfähigkeit

	Prototypizität	
	M	*SD*
Er überprüfte die Schlucht, bevor er sie mit der Canyoning-Gruppe passierte.	3.81	.64
Er setzte seine Idee für die Schülerzeitung mit stichhaltigen Argumenten durch.	3.78	.53
Sie setzte ihren Vorschlag für eine neue Arbeitsaufteilung bei ihrem Chef durch.	3.68	.58
Sie vertrat in einer politischen Diskussion eine klare Meinung und vermochte damit zu überzeugen.	3.68	.71
Er setzte seinen Wunsch, ins Kino zu gehen, gegen die Meinung seiner drei Kollegen durch.	3.62	.59
Sie setzte die Forderung um Ersatz für den kaputten Kochherd beim Vermieter durch.	3.57	.73
Sie verschaffte sich als Trainerin einer Männermannschaft Respekt, indem sie die Mannschaft durch gezieltes Training zum Erfolg führte.	3.50	.61
Sie sprach so lange auf ihren Lehrer ein, bis sie die misslungene Prüfung wiederholen durfte.	3.43	.83
Sie durfte trotz anfänglichem Widerstand des Arztes ihren schwerkranken Freund auf der Intensivstation besuchen.	3.43	.65
Sie brachte ihren Lehrer dazu, einen Fehler einzusehen.	3.41	.69
Sie setzte sich durch klare Kommentare gegen eine Horde wilder Schüler und Schülerinnen durch.	3.41	.69
Er setzte seinen Standpunkt in einer Meinungsverschiedenheit rhetorisch geschickt durch.	3.38	.68
Er beharrte auf der Durchführung der neuen Arbeitsweisung trotz Widerstand seiner Mitarbeiter.	3.30	.85
Sie setzte bei den Ärzten durch, dass ihr Mann die starken Medikamente nicht mehr einnehmen musste, da sie negative Nebenwirkungen befürchtete.	3.24	.89
Sie konnte ihre Freundin überzeugen, dass der mit Unannehmlichkeiten verbundene Vorschlag die Lösung für das Problem war.	3.19	.62
Sie verhinderte eine Erneuerung des Systems, weil sie keinen Vorteil in der Veränderung sah.	3.17	.81
Sie überzeugte ihre Begleiter, welche nach Hause gehen wollten, in der Bar zu bleiben.	3.16	.76
Sie überzeugte den Autofahrer, einen Umweg auf sich zu nehmen, um sie nach Hause zu fahren.	3.16	.73
Er überzeugte die erfahrenen Spieler einer Mannschaft mit guten Leistungen und Selbstvertrauen.	3.08	.76
Sie wehrte sich energisch, als sie bei der Beantwortung einer Frage unterbrochen wurde.	3.05	.74
Er gründete allein eine Jugendquartiergruppe, obwohl der Quartierverein diese für sinnlos hielt.	3.05	.74
Sie setzte ihren Vorschlag für das Kunstwerk in der Gruppe durch, indem sie Leute ausserhalb der Gruppe für sich gewann.	3.05	.70
Er setzte sich gegen die intrigierenden Mitarbeiter durch, indem er die rechtlichen Grundlagen studierte und sich bei seinem Vorgesetzten beklagte.	3.00	.88
Er bestimmte die Getränke- und Speiseauswahl für die Party.	2.97	.90
Er bestimmte die Musik für das Theaterstück.	2.95	.91

Anmerkung. N = 37. Die Skala reicht von 1 = „gar nicht typisch" zu 4 = „sehr typisch".

Anhang 6.11 Prototypenrating der Acts zur Dimension Durchsetzungsfähigkeit (Fortsetzung)

	Prototypizität	
	M	*SD*
Er liess sich durch nichts vom Lernen ablenken.	2.92	.98
Sie verlobte sich gegen den Willen ihrer Eltern.	2.89	.94
Er bestimmte die Mannschaftsaufstellung für das Fussballspiel.	2.86	1.00
Er unterstützte ein Mädchen, welches eine Klasse überspringen wollte, obwohl die Lehrkraft dagegen war.	2.86	.86
Er beharrte darauf, in sein Lieblingsrestaurant zu gehen.	2.84	.83
Er wurde als Stammspieler eingesetzt, weil er doppelt so hart trainierte, nachdem er zuvor in der Mannschaftsaufstellung nicht berücksichtigt worden ist.	2.78	.92
Er bestimmte als Autolenker gegen den Willen der Freunde, frühzeitig nach Hause zu fahren.	2.78	.89
Er ging an die Party, obwohl alle seine Freunde zu Hause blieben.	2.76	.98
Er wählte den Kinofilm aus.	2.76	.95
Er zog gegen den Willen seines Vaters aus, obwohl er von seinen Eltern finanziell abhängig war.	2.76	.89
Sie machte geltend, ihre Anstellung zu kündigen, falls sie die Lohnerhöhung nicht erhielte.	2.73	.93
Sie bestimmte, welches Büromaterial gekauft werden sollte.	2.70	.91
Sie lernte Tag und Nacht, um die Prüfung zu bestehen, obwohl die Erfolgschancen gering waren.	2.65	1.09
Sie änderte den Stundenplan.	2.62	.98
Er hatte bei der Diskussion das letzte Wort, obwohl er im Unrecht war.	2.59	1.04
Er vertrat bei einer politischen Diskussion eine extreme Haltung und beharrte auf seiner Meinung.	2.59	.96
Er erledigte die Arbeit nicht nach Vorschrift, sondern wie er es für richtig empfand.	2.46	.93
Sie kaufte ihrem Sohn den Pullover aus Baumwolle, obwohl dieser den synthetischen mit Aufdruck wollte.	2.24	1.04
Sie gab ihm mit ihrer Körpersprache zu verstehen, dass sie nicht einverstanden war.	2.22	.85
Er weigerte sich, die Schlucht zu besichtigen, weil er die Passfahrt vorzog.	2.22	.79
Sie beendete ihre Partnerschaft, da sie nicht genügend Zeit für sich und ihre Freizeit hatte.	2.16	1.04
Sie reagierte beleidigt, als ihre Freundin nicht so reagierte, wie sie wollte und setzte sich somit durch.	2.14	1.03
Sie manipulierte ihn, indem sie weinte und Hoffnungslosigkeit ausdrückte.	2.00	1.03
Er musste die vom Lehrer verlangte Turnübung nicht durchführen, da er bissige Kommentare äusserte.	1.89	.66
Er drohte seiner Partnerin, die Beziehung zu beenden, falls sie seine Interessen nicht teilte.	1.76	.98
Sie brachte mit ihren offenen Haaren und ihrem tiefen Kleidausschnitt die Männer dazu, ihr zuzuhören.	1.68	.82
Er war den ganzen Abend deprimiert, weil seine Freunde für einmal nicht das unternahmen, was er wollte.	1.59	.80
Er erniedrigte in einer Meinungsverschiedenheit seine Diskussionspartner und machte sich über deren Ansichten lustig.	1.32	.58

Anmerkung. N = 37. Die Skala reicht von 1 = „gar nicht typisch" zu 4 = „sehr typisch".

Anhang 6.12 Prototypenrating der Acts zur Dimension Kontaktfähigkeit

	Prototypizität	
	M	*SD*
Sie nahm die Einladung von ihrem Freund, mit Bekannten von ihm ins Kino zu gehen, ohne zu zögern an und unterhielt sich sofort mit ihnen.	3.86	.35
Er setzte sich am ersten Schultag neben eine Mitschülerin und begann sofort ein Gespräch mit ihr.	3.81	.40
Er begann auf dem Markt in Italien ein Gespräch mit einer Frau und kümmerte sich nicht um die sprachlichen Hindernisse.	3.68	.47
Sie reiste ohne Holländischkenntnisse nach Amsterdam und hatte bereits nach zwei Tagen eine Arbeitsstelle und Freunde gefunden.	3.62	.59
Er begrüsste die Kollegen seines Freundes herzlich und akzeptierte sie schnell als seine eigenen.	3.51	.65
Sie setzte sich in einem beinahe leeren Zug zu einer ihr unbekannten Person und begann ein Gespräch.	3.46	.61
Er sprach auf dem Pausenhof eine fremde Person an und verwickelte sich in ein Gespräch mit ihr.	3.43	.73
Sie war an einer Party innerhalb kurzer Zeit mit einem anderen Gast in ein Gespräch vertieft.	3.41	.60
Sie ging an der Party direkt auf unbekannte Personen zu, welche sympathisch wirkten.	3.41	.69
Sie lebte sich in der neuen Wohngemeinschaft schnell ein.	3.39	.69
Er sprach in der Disco spontan ein Mädchen an.	3.38	.79
Sie übernachtete auf ihrer Reise häufig bei Leuten, die sie unterwegs kennengelernt hatte.	3.38	.72
Sie setzte sich in der Disco an einen Tisch mit fremden Personen und suchte das Gespräch.	3.38	.64
Er begann an der Bushaltestelle ein Gespräch mit einem Mann, während beide auf den Bus warteten.	3.35	.68
Sie integrierte sich hervorragend in den Freundeskreis ihrer Freundin.	3.35	.68
Sie integrierte sich nach ihrem Umzug schnell und mühelos in der neuen Klasse.	3.32	.63
Sie nahm eine fremde Person als Gast auf und erzählte ihr offen von ihrem Leben.	3.30	.81
Er setzte sich im Zug bewusst zu einer Gruppe von Jugendlichen und war trotz des Altersunterschiedes offen für eine Unterhaltung.	3.24	.80
Er lernte im Zug zwei Touristen kennen und lud sie zu sich nach Hause ein.	3.22	.82
Er ging bei der Militäraushebung so auf Unbekannte zu, als würde er sie schon ewig kennen.	3.16	.80
Sie begann während eines Fussballmatches mit zwei unbekannten Männern eine Diskussion über Fussball.	3.14	.75
Sie lernte an einem Abend zehn neue Leute kennen.	3.11	.81
Er duzte an der Party alle Gleichaltrigen und lernte dadurch viele Leute kennen.	2.97	.96
Sie grüsste während eines Stadtspazierganges alle Bekannten herzlich und wusste immer etwas zu erzählen.	2.92	1.02
Sie tauschte im Zug mit einem jungen Mann Adresse und Telefonnummer.	2.86	.89
Er begegnete auf einem Spaziergang zwei ihm unbekannten Frauen und lud sie zu einem Drink ein.	2.84	.96

Anmerkung. *N = 37. Die Skala reicht von 1 = „gar nicht typisch" zu 4 = „sehr typisch".*

Anhang 6.12 Prototypenrating der Acts zur Dimension Kontaktfähigkeit (Fortsetzung)

	Prototypizität	
	M	*SD*
Er half dem neuen Mitglied des Sportvereins, sich zurechtzufinden.	2.81	.81
Er kam beim Einkaufen mit dem Verkäufer ins Gespräch.	2.81	.88
Er lernte an einer Single-Party seine Freundin kennen.	2.69	1.01
Sie entschloss sich, ihre Freundin an eine Party zu begleiten, obwohl sie sich vorgenommen hatte, zu Hause zu bleiben.	2.59	.86
Er gestand seiner Kollegin ohne Mühe, dass er in sie verliebt sei.	2.54	.87
Sie lernte beim Chatten im Internet Leute aus aller Welt kennen.	2.51	1.04
Er ging an der Seepromenade spazieren, weil dort am meisten Leute waren.	2.46	.99
Er bat im Strandbad zwei unbekannte Frauen um etwas Sonnencreme.	2.30	.85
Er ging alleine ins Pub.	2.16	.99
Sie entschloss sich spontan, etwas trinken zu gehen, weil die Schule am Nachmittag ausfiel.	1.76	.72
Er lachte, obwohl er den Witz nicht lustig fand.	1.57	.73

Anmerkung. N = 37. Die Skala reicht von 1 = „gar nicht typisch" zu 4 = „sehr typisch".

Anhang 6.13 Prototypenrating der Acts zur Dimension Verantwortungsbewusstsein

	Prototypizität	
	M	*SD*
Er überprüfte die Schlucht, bevor er sie mit der Canyoning-Gruppe passierte.	3.81	.46
Er konsumierte an der Party keinen Alkohol, da er mit dem Auto unterwegs war.	3.81	.57
Sie liess ihr Auto stehen, weil sie Alkohol getrunken hatte.	3.78	.48
Er lehnte eine interessante Arbeitsstelle im Ausland ab, weil dort die nötige Betreuung für seinen behinderten Sohn nicht gewährleistet gewesen wäre.	3.70	.52
Sie leistete bei einem Unfall erste Hilfe.	3.65	.54
Er kümmerte sich um den Verletzten, welcher regungslos am Waldrand lag.	3.59	.60
Sie half dem verzweifelten Kind, in der Bahnhofshalle seine Mutter wiederzufinden.	3.57	.55
Er nahm seinem betrunkenen Freund die Autoschlüssel weg.	3.57	.55
Er holte seine Kollegin, die sich betrunken hatte, vom Fest ab, fuhr sie nach Hause und kümmerte sich um sie.	3.49	.73
Er besichtigte die Snowboard-Schanze und verbot seiner Gruppe den Sprung.	3.49	.65
Er meldete sich freiwillig bei den Besitzern des Zauns, den er beschädigt hatte.	3.46	.69
Er wies einen Teilnehmer der Wandergruppe, der den Weg verlassen wollte, auf die Gefahren seines Handelns hin.	3.41	.76
Er verhütete mit einem Kondom.	3.41	.76
Er verbrachte viel Zeit mit einer Kollegin, die sich umbringen wollte.	3.38	.68
Er kümmerte sich um seinen kleinen Bruder, das Haus und die Tiere, während seine Eltern in den Ferien weilten.	3.38	.72
Sie half einer unbekannten Person, welche auf offener Strasse tätlich angegriffen wurde.	3.38	.72
Sie schlief nur mit ihm, wenn er ein Kondom benutzte.	3.35	.75
Er meldete sich trotz der ermüdenden Reise wie versprochen unmittelbar nach der Ankunft bei seinem Mitarbeiter.	3.35	.68
Sie räumte nach dem Picknick im Grünen den Abfall weg.	3.35	.72
Er griff in die Schlägerei ein und trennte die Streitenden.	3.27	.87
Sie verzichtete auf ihren freien Abend, da ihre Familie sie brauchte.	3.24	.72
Sie wies die Senioren vor dem Ausflug auf die körperlichen Anforderungen der Wanderung hin.	3.16	.60
Sie setzte sich für eine Bekannte ein, welche zu wenig IV-Rente bekam.	3.14	.79
Sie warnte die Gruppe, die im Wald grillierte, vor der Gefahr eines Feuers.	3.11	.74
Er setzte sich für seinen Mitschüler, der von einer Gruppe belästigt wurde, ein.	3.11	.74
Er versprach die Fertigstellung des Auftrages auf einen Termin, den er sicher einhalten konnte.	3.08	.98
Sie wies ihre Tochter auf die gesundheitsschädigende Wirkung des Rauchens hin.	3.08	.86
Sie fuhr nicht mit ihrem angetrunkenen Kollegen im Auto mit.	3.08	.89
Er informierte die Eltern der Kinder über den Ablauf des Ausfluges der Kindergruppe.	3.05	.78
Sie führte beim Geräteturnen nur ungefährliche Übungen mit den Kindern durch.	3.03	.93

Anmerkung. N = 37. Die Skala reicht von 1 = „gar nicht typisch" zu 4 = „sehr typisch".

Anhang 6.13 Prototypenrating der Acts zur Dimension Verantwortungsbewusstsein (Fortsetzung)

	Prototypizität	
	M	*SD*
Er erledigte seine Hausaufgaben selbstständig und bereitete sich sorgfältig auf die Prüfung vor.	2.97	.96
Er instruierte die Mitglieder des Vereins genau, so dass alle pünktlich erschienen.	2.95	.88
Er überprüfte die Arbeit, um sicherzugehen, dass sie genau und richtig erledigt war.	2.92	.89
Sie führte mehrere Gespräche mit den Mitgliedern des zerstrittenen Teams, um eine vernünftige Lösung zu finden.	2.89	.92
Er fuhr mit dem Unbekannten Skilift, weil er zuvor beobachtet hatte, dass dieser alleine Probleme hatte.	2.86	.82
Er übernahm die gesamte Organisation des Vereinsfestes.	2.84	1.07
Sie übernahm die Aufgabe der Klassenbuchführerin.	2.81	.81
Er liess sich nicht auf den Streit ein und konnte dadurch eine Schlägerei verhindern.	2.81	1.10
Sie riet ihrer Freundin, die über Schmerzen klagte, zum Arzt zu gehen.	2.79	.69
Sie verbesserte die Arbeiten ihrer Teammitglieder, damit das Gruppen-Projekt angenommen wurde.	2.78	.92
Sie kontrollierte zweimal, ob alle Geräte ausgeschaltet waren, bevor sie die Wohnung verliess.	2.78	.85
Er trennte den Abfall sorgfältig.	2.78	.82
Sie arbeitete stundenlang und half danach noch ihren Arbeitskollegen, bis alles erledigt war.	2.78	.85
Er leitete ein Ferienlager mit vielen jüngeren Kindern.	2.76	.98
Sie machte eine Aufstellung ihrer Rechnungen, um alles rechtzeitig zu bezahlen.	2.76	1.01
Sie organisierte den Abschlussball.	2.72	.94
Er trug während eines langen Marsches im Militärdienst zusätzlich das Gepäck eines Kollegen.	2.70	.97
Sie spendete Blut.	2.68	.88
Sie hatte ein schlechtes Gewissen, weil sie zu spät zur Verabredung erschien.	2.68	1.00
Er holte sich für die Klassenreise Offerten von Reisebüros ein.	2.62	.83
Sie erschien zur Prüfung, obwohl sie krank war.	2.57	.96
Sie holte ihre Tochter nachts um drei Uhr bei ihrem Freund ab, da sich diese gestritten hatten.	2.43	.96
Er regelte die Finanzen für das Ferienlager.	2.41	1.01
Sie liess ihre Freundin nicht mit einem unbekannten Mann ausgehen.	2.41	.96
Sie spendete Geld für die Schulausbildung eines pakistanischen Mädchens.	2.35	.82
Er übernahm die Verantwortung für die Homepage des Badmintonklubs.	2.35	.82
Er verbot seiner Tochter, eine Cola zu trinken, da das darin enthaltene Koffein schädlich ist.	2.32	.67
Sie kaufte Bio-Fleisch, da sie kein Fleisch mit Antibiotika essen wollte.	2.24	.89
Sie seilte die Kursteilnehmer über eine Felswand ab.	2.08	1.05
Sie übernahm die Verantwortung für ihre Kollegin, die bei der Prüfung von ihr abschrieb.	1.81	.84

Anmerkung. N = 37. Die Skala reicht von 1 = „gar nicht typisch" zu 4 = „sehr typisch".

Anhang 6.14 Faktorladungen der Items des Leadership-Fragebogens (Version mit 39 Items; Datensatz 2003; *N* = 7'871)

		Faktor 1	Faktor 2	Faktor 3
Kontaktfähigkeit	Zugfahrt	**.69**	.16	.01
	Flugzeug	**.62**	.23	-.03
	Begleitung	**.61**	.11	.05
	Fitness	**.60**	.14	-.04
	Geburtstag	**.60**	.15	.00
	Barmann	**.58**	.03	.01
	Kurs	**.56**	.28	-.03
	Schultag	**.54**	.18	.03
	Nachbarn	**.53**	**.35**	-.04
	Alleingelassen	**.52**	.14	.08
	Party	**.50**	.07	.08
	Umzug	**.49**	**.31**	-.04
	Zelten	**.46**	.03	.06
Verantwortungsbewusstsein	Ampel	.22	**.54**	-.07
	Bergtour	.12	**.51**	.08
	Silvester	.19	**.50**	-.09
	Autofahren	.14	**.50**	.08
	Nachhilfestunden	.20	**.49**	-.02
	Wohnung	.08	**.48**	.17
	Schanze	.00	**.47**	.00
	Beratungsstelle	.15	**.45**	.04
	Kind	.18	**.45**	-.04
	Subventionen	.27	**.40**	-.02
	Unstimmigkeiten	.27	**.38**	.03
	Mädchen	.25	**.34**	-.09
	Malediven	.03	**.34**	.15
Durchsetzungsfähigkeit	Aufräumen	-.03	.12	**.46**
	Fahrer	-.10	-.04	**.45**
	Waschküche	.03	.14	**.45**
	Unterbruch	.02	.06	**.44**
	Geschirr	.03	.29	**.40**
	Arbeit	.00	.03	**.40**
	Lohnerhöhung	.13	.09	**.40**
	Auswärts	-.07	-.07	**.39**
	Musik	-.07	-.20	**.38**
	Probleme	-.06	-.16	**.38**
	Disco	.06	-.28	**.38**
	Schülerzeitung	.14	.17	**.33**
	Zugreise	.09	.02	**.26**
Eigenwert		6.37	2.19	1.76
erklärte Varianz (%)		11.91	8.95	5.60

Anhang 6.15 Faktorladungen der Items des Leadership-Fragebogens (Version mit 30 Items; Datensatz 2003; *N* = 7'871)

		Faktor 1	Faktor 2	Faktor 3
Kontaktfähigkeit	Zugfahrt	**.70**	.15	.00
	Flugzeug	**.65**	.22	-.03
	Fitness	**.63**	.12	-.03
	Begleitung	**.63**	.10	.05
	Geburtstag	**.60**	.13	.00
	Barmann	**.60**	.01	.02
	Kurs	**.58**	.27	-.02
	Schultag	**.57**	.17	.04
	Nachbarn	**.55**	**.34**	-.04
	Umzug	**.49**	**.30**	-.04
Verantwortungsbewusstsein	Ampel	.23	**.54**	-.05
	Autofahren	.13	**.53**	.08
	Bergtour	.12	**.52**	.09
	Silvester	.19	**.51**	-.09
	Nachhilfestunden	.19	**.51**	-.04
	Schanze	-.01	**.50**	-.01
	Wohnung	.09	**.48**	.18
	Beratungsstelle	.15	**.46**	.04
	Kind	.18	**.46**	-.05
	Subventionen	.28	**.40**	-.04
Durchsetzungsfähigkeit	Fahrer	-.08	-.05	**.48**
	Waschküche	.03	.13	**.47**
	Aufräumen	-.03	.13	**.47**
	Unterbruch	.02	.06	**.44**
	Auswärts	-.05	-.10	**.44**
	Arbeit	.03	.01	**.44**
	Geschirr	.04	.29	**.43**
	Probleme	-.04	-.18	**.41**
	Musik	-.07	-.21	**.39**
	Lohnerhöhung	.13	.08	**.39**
Eigenwert		5.43	2.01	1.57
erklärte Varianz (%)		13.23	10.17	6.59

Anhang 6.16 Normierung des Leadership-Fragebogens (Version mit 30 Items; Datensatz 2003; *N* = 7'871)

Durchsetzungsfähigkeit					Kontaktfähigkeit					Verantwortungsbewusstsein				
Score	*n*	%	kum		Score	*n*	%	kum		Score	*n*	%	kum	
10	2	.0	.0		10	17	.2	.2		10	3	.0	.0	
11	4	.1	.1		11	17	.2	.4		11	21	.3	.3	
12	5	.1	.1		12	23	.3	.7		12	12	.2	.5	
13	12	.2	.3	1	13	34	.4	1.2	1	13	16	.2	.7	
14	22	.3	.6		14	32	.4	1.6		14	19	.2	.9	1
15	47	.6	1.2		15	52	.7	2.2		15	30	.4	1.3	
16	79	1.0	2.2		16	69	.9	3.1		16	45	.6	1.9	
17	173	2.2	4.4	2	17	98	1.2	4.3		17	54	.7	2.5	
18	242	3.1	7.4		18	159	2.0	6.4	2	18	88	1.1	3.7	
19	395	5.0	12.5	3	19	243	3.1	9.5		19	108	1.4	5.0	
20	597	7.6	20.0		20	331	4.2	13.7	3	20	125	1.6	6.6	2
21	741	9.4	29.5	4	21	374	4.8	18.4		21	225	2.9	9.5	
22	826	10.5	40.0		22	446	5.7	24.1		22	232	2.9	12.4	
23	921	11.7	51.7	5	23	474	6.0	30.1	4	23	314	4.0	16.4	3
24	878	11.2	62.8		24	502	6.4	36.5		24	421	5.3	21.8	
25	785	10.0	72.8	6	25	539	6.8	43.3		25	515	6.5	28.3	4
26	680	8.6	81.4		26	519	6.6	49.9	5	26	615	7.8	36.1	
27	497	6.3	87.7	7	27	544	6.9	56.8		27	722	9.2	45.3	5
28	374	4.8	92.5		28	521	6.6	63.4		28	831	10.6	55.9	
29	228	2.9	95.4	8	29	551	7.0	70.4	6	29	777	9.9	65.7	6
30	157	2.0	97.4		30	548	7.0	77.4		30	757	9.6	75.3	
31	99	1.3	98.6		31	441	5.6	83.0	7	31	651	8.3	83.6	7
32	44	.6	99.2		32	418	5.3	88.3		32	513	6.5	90.1	
33	21	.3	99.5		33	340	4.3	92.6	8	33	378	4.8	94.9	8
34	16	.2	99.7		34	249	3.2	95.8		34	200	2.5	97.5	
35	13	.2	99.8	9	35	177	2.2	98.1		35	119	1.5	99.0	
36	7	.1	99.9		36	86	1.1	99.1		36	56	.7	99.7	
37	4	.1	100.0		37	43	.5	99.7	9	37	20	.3	99.9	9
38	1	.0	100.0		38	20	.3	99.9		38	4	.1	100.0	
39	1	.0	100.0		39	1	.0	100.0		39	0	.0	100.0	
40	0	.0	100.0		40	3	.0	100.0		40	0	.0	100.0	

7. Überprüfung des Leadership-Fragebogens

7.1 Reanalyse der gekürzten Version des Leadership-Fragebogens

Die vorgenommene Reduktion der Skalen des Leadership-Fragebogens von 13 auf 10 Items überprüfe ich anhand des Datensatzes aus dem Jahr 2008, welcher die Testergebnisse von insgesamt 21'008 deutschsprachigen Stellungspflichtigen umfasst. Da ich in den vergangenen Jahren die Erfahrung gemacht habe, dass es unter den Stellungspflichtigen eine kleine Minderheit gibt, welche einzelne Testverfahren unseriös ausfüllt oder von der Bearbeitung eines sehr sprachlastigen Tests überfordert ist, unterziehe ich den Datensatz einer umfassenderen Bereinigung als den Datensatz aus dem Jahre 2003. Ich führe dazu folgende drei Schritte durch, wobei die Bestimmung der Ausreisser jeweils über Box-Plots erfolgt:

1. Testbearbeitungszeit: Unter der Annahme, dass einzelne Stellungspflichtige den Test entweder unseriös ausgefüllt haben oder Probleme mit dem Sprachverständnis hatten, schliesse ich diejenigen mit einer sehr kurzen (< 262 sec; n = 127, 0.6%) oder sehr langen (> 1'418 sec; n = 347, 1.6%) Testbearbeitungszeit aus dem Datensatz aus.

2. Durchklicken, Variante 1: Die Durchsicht der Daten ergab, dass bei einigen Datensätzen jeweils nur eine Antwortalternative eingegeben wurde. Wahrscheinlich handelt es sich hierbei um Fälle, bei welchen die Testassistenten den Test aufgeschaltet haben, der Stellungspflichtige jedoch nicht zur Bearbeitung des Testblocks erschienen ist. Die Testassistenten haben dann den Test selbst durchgeklickt, um den Testblock für den betreffenden Stellungspflichtigen zu beenden. Es könnte jedoch auch zutreffen, dass vereinzelt Stellungspflichtige den Test auf Grund ihrer Motivationslage unseriös bearbeitet und ihn einfach durchgeklickt haben, indem sie jeweils mehrmals hintereinander dieselbe Antwortkategorie wählten. Die Kontrolle des sehr häufigen oder sehr seltenen Auftretens einer Antwortkategorie bei einem Stellungspflichtigen mittels Box-Plots führt zum Ausschluss von insgesamt 561 (2.7%) Datensätzen, wobei ich folgende Kriterien bestimmte:
   Antwortalternative 1: > 15 oder null Mal gewählt (n = 160; n = 1)
   Antwortalternative 2: > 13 oder ein Mal gewählt (n = 249; n = 27)
   Antwortalternative 3: > 13 oder ein Mal gewählt (n = 48; n = 71)
   Antwortalternative 4: > 14 Mal gewählt (n = 50)

3. Durchklicken, Variante 2: Um Stellungspflichtige aus dem Datensatz auszuschliessen, welche unsystematisch und wahrscheinlich unreflek-

tiert eine Antwort gewählt haben, bestimme ich für jeden Stellungspflichtigen die Streuung der Antworten (Scores) pro Dimension des Leadership-Fragebogens. Total 172 (0.9%) Stellungspflichtige schliesse ich auf Grund einer auffallend grossen Streuung aus:
Durchsetzungsfähigkeit: Streuung grösser als 1.35 (n = 70)
Kontaktfähigkeit: Streuung grösser als 1.26 (n = 44)
Verantwortungsbewusstsein: Streuung grösser als 1.33 (n = 59)

Insgesamt habe ich somit 1'207 (5.7%) Datensätze ausgeschlossen, was zu einem Gesamtdatensatz führt, welcher die Testresultate von 19'801 Stellungspflichtigen beinhaltet. Für die Angabe des durchschnittlichen Alters dieser Stellungspflichtigen muss ich auf die Jahresstatistik der Rekrutierung (Kommando Rekrutierung, 2009) zurückgreifen, da mir aus Gründen des Datenschutzes nur ein pseudonymisierter Datensatz zur Verfügung steht, in welchem die Altersangabe fehlt. Das durchschnittliche Alter der Stellungspflichtigen lag 2008 bei knapp 20 Jahren (zwischen 19.8 und 19.9 Jahren) und der Frauenanteil betrug 0.4%. In Tabelle 7.1 ist die Aufteilung der Stichprobe auf die sechs Rekrutierungszentren dargestellt. Der Vergleich mit der Rekrutierungsstatistik zeigt, dass – ausgehend vom Gesamtdatensatz (N = 21'008) – für das Jahr 2008 die Leadership-Fragebogen-Daten von ca. 75% aller deutschschweizer Stellungspflichtigen vorliegen. Diese Differenz lässt sich zu einem grossen Teil damit erklären, dass einige der als dienstuntauglich beurteilten Stellungspflichtigen nicht mehr zum zweiten Psychologie-Testblock erscheinen, welcher auch den Leadership-Fragebogen umfasst. Weiter gibt es Stellungspflichtige, die schon einmal an einer Rekrutierung teilgenommen haben und dort aus medizinischen Gründen zu einer Nachbeurteilung zurückgestellt wurden und die beim erneuten Erscheinen im Rekrutierungszentrum keine Tests mehr zu absolvieren haben.

Tabelle 7.1
Aufteilung der Stichprobe 2008 nach Rekrutierungszentrum

	Stichprobe 2008		Rekrutierungsstatistik 2008	
Rekrutierungszentrum	*n*	%	*n*	%
Lausanne (französischsprachig)	1	.01	–	–
Sumiswald	5'097	25.74	6'075	21.85
Monte Ceneri (italienischsprachig)	2	.01	–	–
Windisch	6'525	32.95	9'140	32.87
Rüti	4'506	22.76	7'479	26.90
Mels	3'670	18.53	5'114	18.39
Total	19'801	100.00	27'808	100.01

In Abbildung 7.1 ist der Vergleich der anhand der Datensätze 2008 und 2003 berechneten Rohwert-Verteilungen der drei Leadership-Skalen dargestellt. Es zeigt sich sowohl anhand der Grafiken wie auch der Kennzahlen, dass sich die jeweiligen Verteilungs-Paare kaum voneinander unterscheiden. Auch der Vergleich der anhand der Datensätze von 2003 und 2008 berechneten Item- und Skalenkennwerte ergibt keine nennenswerten Unterschiede. Die Tabellen mit den Angaben zu den einzelnen Items und die Korrelationsmatrix aller Items des Fragebogens sind in den Anhängen 7.1 und 7.2 dargestellt. In Tabelle 7.2 habe ich die Skalenkennwerte, in Tabelle 7.3 die Korrelationsmatrix der Leadership-Skalen und die durchschnittlichen Item-Interkorrelationen aufgeführt.

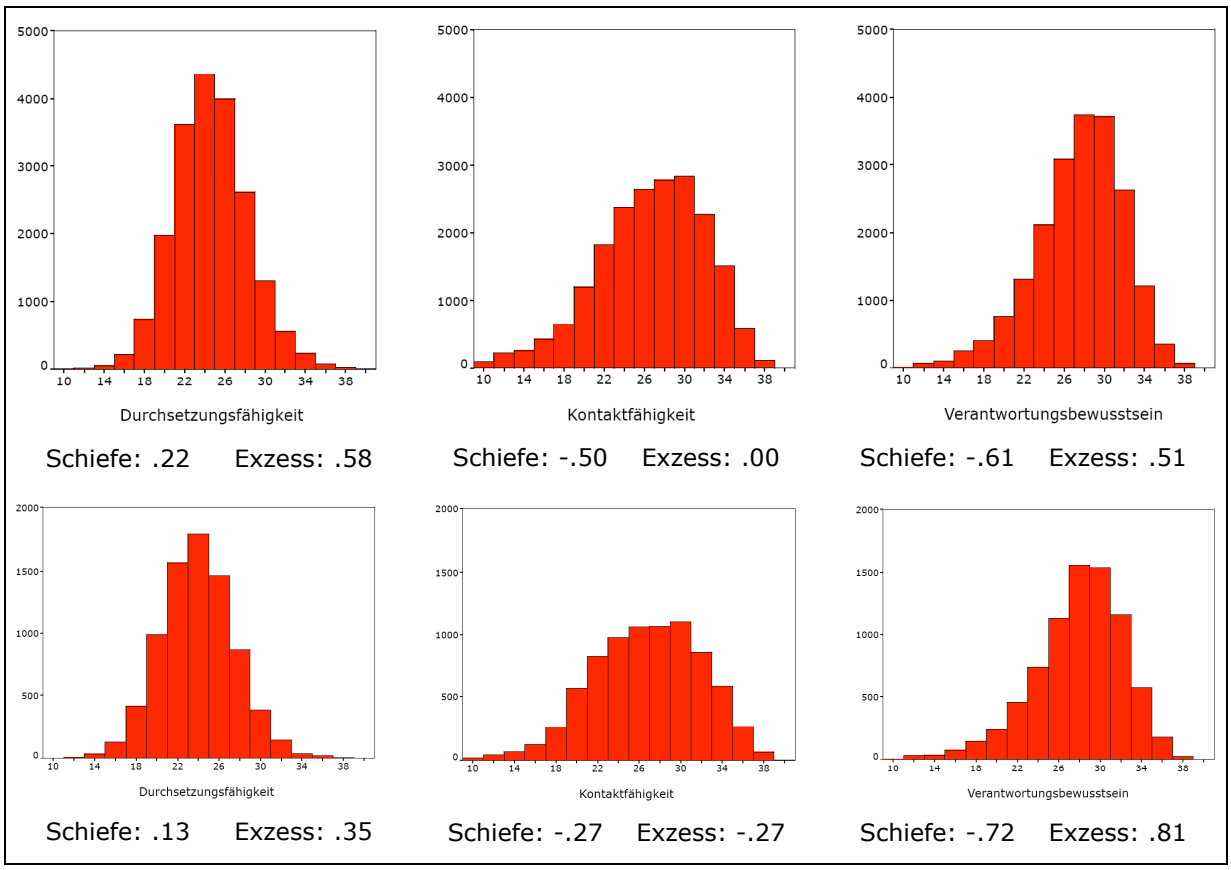


Abbildung 7.1 Vergleich der Rohwert-Verteilungen der drei Skalen des Leadership-Fragebogens: Datensätze 2008 (*N* = 19'801) und 2003 (*N* = 7'871).

Tabelle 7.2
Vergleich der Datensätze 2008 und 2003: Skalenkennwerte

	Datensatz 2008 (N = 19'801)			Datensatz 2003 (N = 7'871)		
Skala	*M*	*SD*	α	*M*	*SD*	α
Durchsetzungsfähigkeit	24.13	3.69	.57	23.43	3.55	.53
Kontaktfähigkeit	26.18	5.35	.84	26.30	5.14	.83
Verantwortungsbewusstsein	27.02	4.37	.70	27.45	4.31	.71

Tabelle 7.3
Vergleich der Datensätze 2008 und 2003: Korrelationsmatrix

	Datensatz 2008 (N = 19'801)			Datensatz 2003 (N = 7'871)		
Skala	DF	KF	VB	DF	KF	VB
Durchsetzungsfähigkeit	*.12*			*.10*		
Kontaktfähigkeit	-.08	*.34*		-.00	*.33*	
Verantwortungsbewusstsein	-.03	.53	*.20*	.04	.53	*.20*

Anmerkung. Alle Korrelationen mit Ausnahme der Nullkorrelation sind hochsignifikant. Kursiv in der Diagonale sind die durchschnittlichen Item-Interkorrelationen aufgeführt.

Zur erneuten Bestimmung der Faktorenstruktur wiederhole ich mit den Daten von 2008 zu den 30 Items des Leadership-Fragebogens eine Hauptkomponentenanalyse mit orthogonaler Rotation (Varimax). Die Voraussetzungen für die Durchführung einer Faktorenanalyse sind gegeben: Das Mass der Stichprobeneignung nach Kaiser-Meyer-Olkin beträgt *KMO* = .93, was nach Kaiser (1974) als ‚fabelhaft' zu beurteilen ist. Zudem sind alle KMO-Werte der einzelnen Variablen grösser als .76 und liegen somit deutlich über der Grenze von .50. Auch der Bartlett-Test auf Sphärizität wird signifikant ($\chi^2(435, N = 19'801) = 90'652.25$, $p < .001$), was auf genügend hohe Korrelationen zwischen den Items hinweist. Die Determinante der Korrelationsmatrix beträgt $|R| = 0.01025$ und liegt somit über dem Grenzwert von .00001 (Field, 2009). Jedoch wird der Test nach Haitovsky (1969, siehe auch Rockwell, 1975) nicht signifikant ($\chi^2_H(435, N = 19'801) = 203.89$, $p > .05$), was ein Anzeichen auf Multikollinearität sein könnte. Zur Bestimmung der Anzahl Faktoren verwende ich wiederum den Scree-Plot. Er zeigt klar drei Faktoren auf (siehe Abbildung 7.2), welche zusammen 30.78% der Varianz erklären. In Tabelle 7.4 sind die Faktorladungen nach der Rotation aufgeführt.

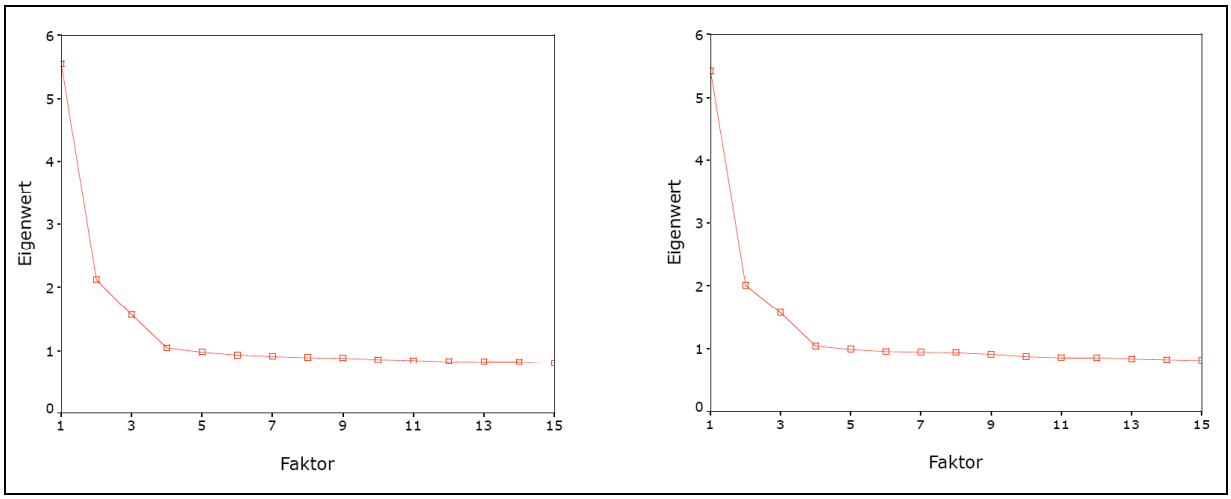


Abbildung 7.2 Scree-Plots der Faktorenanalysen der 30-Item-Versionen des Leadership-Fragebogens (Datensätze 2008 und 2003).

Alle oben beschriebenen Auswertungen zeigen, dass es zwischen den Datensätzen 2003 und 2008 keine bedeutsamen Unterschiede gibt. Wiederum zeigt sich in der explorativen Faktorenanalyse, dass die einzelnen Items auf die intendierten Skalen laden und nur zwei Items der Skala Kontaktfähigkeit eine Nebenladung $\geq$ .30 auf einen anderen Faktor, Verantwortungsbewusstsein, aufweisen. Zudem konnte ich auch die 2003 bestimmten Reliabilitäten der drei Skalen bestätigen, wobei diejenige der Skala Durchsetzungsfähigkeit erneut als ungenügend zu taxieren ist (Evers, 2001; siehe auch Kapitel 6.4). Ob dies an ungeeigneten Items oder an einer grundsätzlichen Schwierigkeit der Operationalisierung dieser Verhaltensdimension liegt, kann ich auf Grund der Daten nicht beantworten. Immerhin konnte ich aufzeigen, dass das gewählte Wertequadrat wie gewünscht funktioniert. Denkbar ist jedoch, dass sich gerade bei der Durchsetzungsfähigkeit ein Problem der anlässlich der Rekrutierung stark ausgeprägten Selbstdarstellungstendenzen zeigt, da es für die Stellungspflichtigen nicht eindeutig klar ist, wo sich auf dieser Dimension das Optimum befindet: Zeigt ein idealer militärischer Vorgesetzter eine hohe Durchsetzungsfähigkeit oder eher eine mittelstark ausgeprägte? Als naheliegende Möglichkeit, die Reliabilität zu verbessern, böte sich eine Verlängerung der Skala an: Um eine Reliabilität von r'_{tt} = .70 zu erreichen, müsste ich die Skala um acht Items verlängern, für r'_{tt} = .80 um 20 (Formel nach Lienert & Raatz, 1998). Eine solche Verlängerung kommt jedoch nicht in Frage, da die Bearbeitung des Fragebogens dann zu viel Zeit in Anspruch nehmen würde und für viele Stellungspflichtige zu anstrengend wäre. Als letzte Möglichkeit einer Verbesserung der Reliabilitäten der Skalen sehe

ich noch den Einsatz alternativer Scoring-Methoden, welche ich im nächsten Kapitel darstellen werde.

Tabelle 7.4
Faktorladungen der Items des Leadership-Fragebogens (Varimaxrotation; Datensatz 2008; N = 19'801)

		Faktor 1	Faktor 2	Faktor 3
Kontaktfähigkeit	Zugfahrt	**.69**	.22	-.06
	Begleitung	**.65**	.12	.03
	Flugzeug	**.65**	.22	-.06
	Fitness	**.63**	.12	-.04
	Kurs	**.61**	.26	-.03
	Geburtstag	**.61**	.13	.02
	Barmann	**.60**	-.01	.01
	Schultag	**.59**	.16	.03
	Nachbarn	**.56**	**.34**	-.07
	Umzug	**.53**	**.30**	-.07
Verantwortungsbewusstsein	Ampel	.22	**.55**	-.08
	Bergtour	.09	**.55**	.06
	Silvester	.19	**.52**	-.09
	Autofahren	.17	**.51**	.07
	Schanze	.01	**.50**	-.03
	Wohnung	.02	**.49**	.17
	Kind	.20	**.48**	-.04
	Beratungsstelle	.18	**.47**	.01
	Nachhilfestunden	.24	**.45**	-.05
	Subventionen	.23	**.40**	-.05
Durchsetzungsfähigkeit	Aufräumen	.00	.03	**.51**
	Waschküche	.03	.07	**.48**
	Unterbruch	.02	.04	**.47**
	Fahrer	-.09	-.02	**.47**
	Arbeit	-.02	.04	**.45**
	Geschirr	.04	.23	**.45**
	Musik	-.05	-.20	**.43**
	Auswärts	-.13	-.05	**.42**
	Probleme	-.07	-.17	**.41**
	Lohnerhöhung	.13	.00	**.41**
Eigenwert		5.55	2.12	1.56
erklärte Varianz (%)		18.51	7.06	5.21

7.2 Auswirkungen unterschiedlicher Scoring-Arten auf die Reliabilität der Leadership-Skalen

Alle bisher durchgeführten Analysen der verschiedenen mir zur Verfügung stehenden Datensätze haben gezeigt, dass die Reliabilität der Skala Durchsetzungsfähigkeit ungenügend ist. Es stellt sich nun die Frage, ob sich diese verbessern lässt, indem ich für die vorgegebenen vier Antwortalternativen andere Gewichtungen einführe. Bisher habe ich jeweils die für Persönlichkeits-Fragebogen übliche Skalierung eingesetzt, bei welcher die zugewiesenen Punktewerte die Ausprägung der jeweiligen Antwortalternative innerhalb der erfassten Dimension widerspiegeln. So weise ich zum Beispiel bei der Skala Durchsetzungsfähigkeit der Übertreibung der erwünschten Antwortalternative – also der Rücksichtslosigkeit – den höchsten Punktewert zu, da diese auch der höchsten Ausprägung auf dieser Persönlichkeitsdimension entspricht. Dem Wertequadrat liegt jedoch der Gedanke zugrunde, dass die zwei Tugenden als gleichwertig aufzufassen sind und die jeweiligen Übertreibungen unerwünschte Auswüchse der Tugenden sind (z. B. Gloor, 2007): In der Logik des Wertequadrates entspricht das Maximum nicht dem Optimum. Dies gilt im Grund genommen auch bei den nach dem üblichen Vorgehen entwickelten Persönlichkeitsskalen, wenn Personalverantwortliche diese im Selektionskontext einsetzen: Bei vielen Persönlichkeitsdimensionen – als leicht nachvollziehbares Beispiel sei hier die Gewissenhaftigkeit genannt – entsprechen die Höchstwerte häufig nicht der gemäss dem Anforderungsprofil bestimmten idealen Ausprägung. In einem die Gewichtung eines Wertequadrats abbildenden Scoring müsste ich also den Übertreibungen tiefere Werte zuweisen als den Tugenden. In Abbildung 7.3 sind die bisher eingesetzte Skalierung und eine mögliche, den Grundgedanken des Wertequadrats übernehmende, Alternative dazu dargestellt.

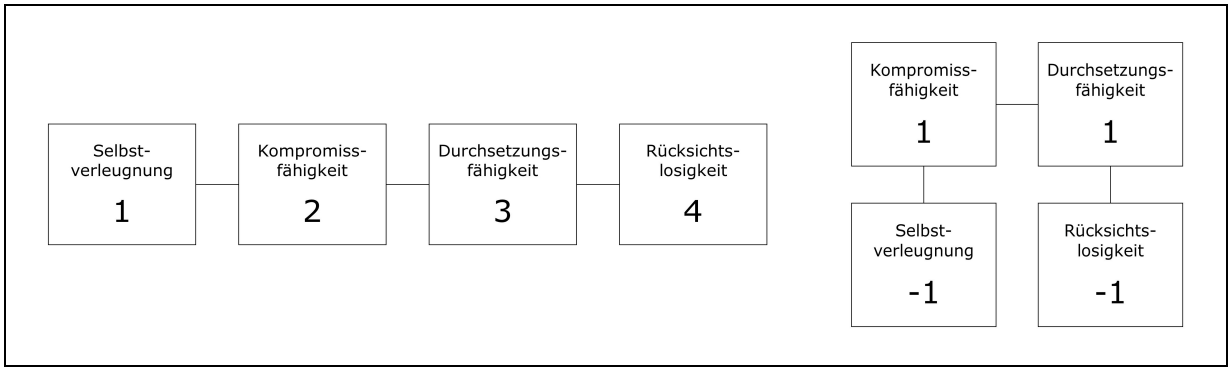


Abbildung 7.3 Zwei mögliche Scoring-Varianten von Wertequadrat-Items.

In Tabelle 7.5 sind die Auswirkungen verschiedener Scoring-Arten auf die Reliabilität der Skalen aufgeführt. An dieser Stelle könnte der Einwand erfolgen, dass schlussendlich nur die Validität für die Güte eines Tests entscheidend ist und somit der Reliabilität eine nicht allzu grosse Bedeutung beizumessen ist. Rost (1996) zeigt in seiner Diskussion der logischen Beziehungen zwischen den drei Hauptgütekriterien von Testverfahren auf, dass ein Test mit einer geringen Reliabilität keine hohe externe Validität erzielen kann. Er weist jedoch auch auf das Reliabilitäts-Validitäts-Dilemma (*attenuation paradox*; Loevinger, 1954) hin, das auftritt, wenn ein sehr homogener Test, welcher eine eng umschriebene Persönlichkeitsdimension erfasst, nur noch eine geringe Korrelation zum Validitätskriterium aufweist. Mit dem Einsatz des AFA laufen wir beim Leadership-Fragebogen jedoch nicht Gefahr, einen einseitig homogenen Test zu entwickeln, da wir das zu erfassende Konstrukt durch eine breite Auswahl an Verhaltensweisen in verschiedenen Situationen operationalisieren. So lässt sich davon ausgehen, dass hochreliable Skalen hier keinen negativen Einfluss auf die Kriteriumsvalidität ausüben.

Tabelle 7.5

Auswirkungen verschiedener Scoring-Varianten auf die Reliabilität (Cronbach Alpha) der drei Leadership-Skalen

Scoring-Variante	Zuordnung der Scoring-Werte zu den vier Wertequadranten der drei Wertequadrate				Reliabilität der Leadership-Skalen		
	Selbst-verleugnung	Kompromiss-fähigkeit	Durchsetzungs-fähigkeit	Rücksichts-losigkeit	Durchsetzungs-fähigkeit	Kontaktfähigkeit	Verantwortungs-bewusstsein
	Menschenscheu	Zurückhaltung	Kontaktfähigkeit	Distanzlosigkeit			
	Verantwor-tungslosigkeit	Bewusstsein für Eigenverantw.	Verantwortungs-bewusstsein	Bevormundung			
LS_1	1	2	3	4	.57	.84	.70
LS_2	1	3	4	5	.55	.84	.72
LS_3	-1	2	2	3	.49	.78	.67
WQ_1	-1	1	1	-1	.41	.57	.42
WQ_2	-1	0	1	-1	.30	.57	.40
WQ_3	-1	1	1	0	.42	.68	.52

Anmerkung. N = 19'801.

Bei der in oben stehender Tabelle aufgeführten alternativen Scoring-Varianten habe ich neben der Wertequadrat-Logik auch das Anforderungsprofil berücksichtigt: So ist das völlige Fehlen der erwünschten Verhaltensweise – zum Beispiel Verantwortungslosigkeit – bei militärischem Kader eine gravierendere Abweichung vom Idealprofil, als deren Übertreibung – also Bevormundung. Die

durchgeführten Reliabilitätsberechnungen zeigen jedoch auf, dass keine der fünf simulierten alternativen Scoring-Varianten zu einer substanziellen Erhöhung des Cronbach Alphas führt. Im Gegenteil: Die Werte fallen vor allem bei den Wertequadrat-Skalierungen (WQ_1 bis WQ_3) deutlich tiefer aus als bei der ursprünglichen Scoring-Variante. Dies könnte damit zusammenhängen, dass wir während des gesamten Itemselektion-Prozesses das herkömmliche Scoring eingesetzt haben. Die Ursache für diesen Effekt könnte aber auch darauf zurückzuführen sein, dass uns die Operationalisierung der Wertequadrate nicht gelungen ist. Dies habe ich anhand des Datensatzes aus dem Jahre 2008 auf der Grundlage folgender – dem Konzept der Distraktorenanalyse (z. B. Lienert & Raatz, 1998) angelehnten – Überlegung überprüft: Wenn ein Stellungspflichtiger beim Item X der Skala Y die Antwortalternative mit dem tiefsten Scoring-Wert wählt, so muss die Summe aller anderen von ihm gewählten Antworten in der Skala Y tiefer sein, als bei Stellungspflichtigen, welche eine der drei anderen Antwortalternativen dieses Items gewählt haben. Ich habe also bei allen Items des Leadership-Fragebogens für jede Antwortalternative den Mittelwert der Summe der Antwort-Scores der Items der betreffenden Skala der jeweiligen Wählergruppe berechnet. Ist die Operationalisierung des Wertequadrates geglückt, so müssen sich die vier Mittelwerte pro Item entsprechend den Scoring-Werten der einzelnen Antwortalternativen in aufsteigender Reihenfolge präsentieren. Dies trifft auch – wie in den Tabellen in den Anhängen 7.3 bis 7.5 ersichtlich – ausnahmslos für alle Items zu. Zur Veranschaulichung der durchgeführten Berechnungen stelle ich in der nachfolgenden Tabelle die Werte von zwei Items der Skala Kontaktfähigkeit dar. Es ist ersichtlich, dass sich beim Item „Begleitung" die Mittelwerte der vier Antwortalternativen deutlich voneinander unterscheiden, wohingegen diese beim Item „Nachbarn" bei den Antwortalternativen 3 und 4 beinahe gleich hoch ausfallen.

Tabelle 7.6

Ausschnitt aus der Tabelle mit den Werten zur Überprüfung der Wertequadrate der Skala Kontaktfähigkeit

	r_{it}		*n*	%	min	max	*M*	*SD*
Begleitung	.62	WQ 1	1'409	7.12	10	33	17.23	5.04
		WQ 2	6'657	33.62	11	37	23.14	3.91
		WQ 3	10'045	50.73	12	39	28.45	3.53
		WQ 4	1'690	8.53	16	40	32.13	3.33
Nachbarn	.44	WQ 1	3'881	19.60	10	35	21.75	5.60
		WQ 2	7'347	37.10	11	37	24.45	4.14
		WQ 3	6'736	34.02	13	38	29.61	3.58
		WQ 4	1'837	9.28	13	40	29.89	4.08

Somit bleibt das Problem mit der ungenügend hohen Reliabilität der Skala Durchsetzungsfähigkeit bestehen: Eine Verlängerung der Skala um bis zu 20 Items ist nicht umsetzbar und eine alternative Scoring-Variante führt zu einer noch tieferen Reliabilität. Eine weitere – mit Bestimmtheit Erfolg versprechende – Möglichkeit zur Reliabilitätssteigerung ist die Abkehr vom Forced-Choice-Antwortformat und die Einführung einer likert-skalierten Einstufung jeder der vier Verhaltensweisen. Dies führt de facto zu einer Vervierfachung der Itemanzahl und somit zwangsläufig zu einer höheren Reliabilität. Zudem lassen sich zwei Nachteile des Forced-Choice-Formates überwinden: Neu habe ich auch Informationen von den ursprünglich nicht gewählten Antwortalternativen und die Testbearbeiter können genau angeben, wie stark die geschilderte Verhaltensweise mit ihrem tatsächlich gezeigten Verhalten übereinstimmt. Gleichzeitig komme ich mit der Einführung der Möglichkeit, jede Antwortalternative einstufen zu können, auch der ab und zu von Stellungspflichtigen geäusserten Bemerkung nach, dass sie in den geschilderten Situationen der Dimension Durchsetzungsfähigkeit durchaus unterschiedliche Reaktionen zeigen würden: Wenn eine moderate Verhaltensweise nicht zum Ziel führt, wenden sie eine dominantere an.

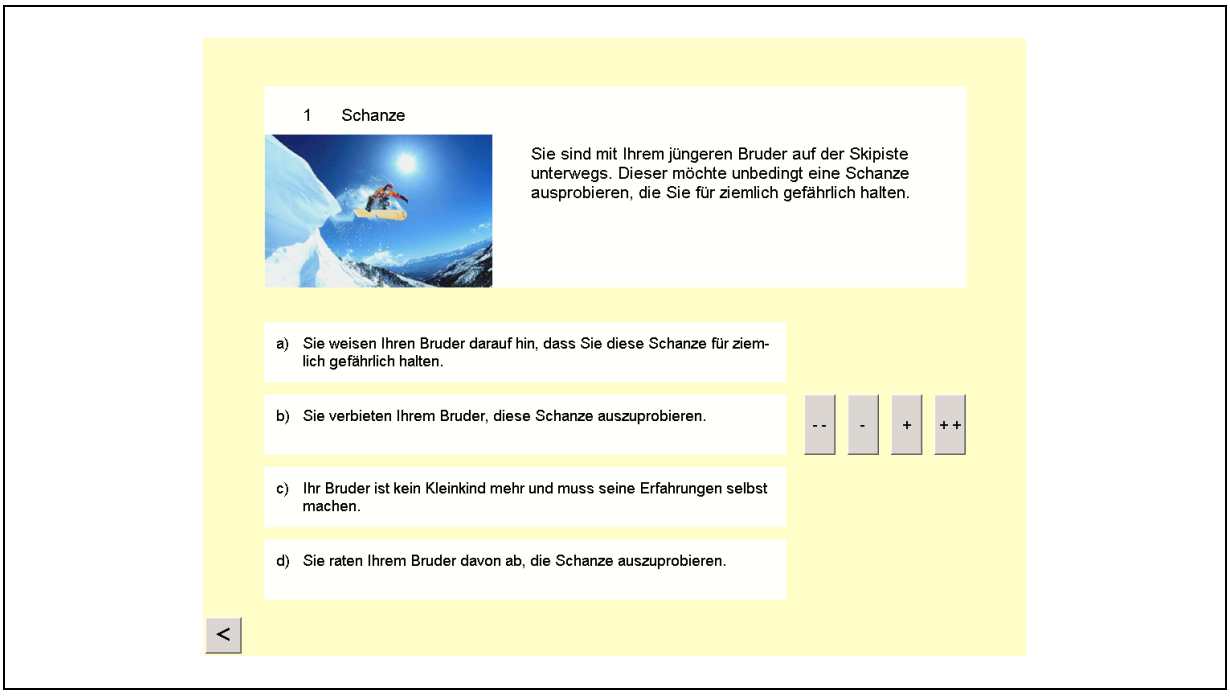


Abbildung 7.4 Itemlayout der likert-skalierten Version des Leadership-Fragebogens.

Um eine Version des Leadership-Fragebogens mit likert-skalierten Antwortformat zu überprüfen, programmierten wir diese – wie in Abbildung 7.4 dargestellt mit

einer vierstufigen Antwortskala – in das Testsystem der Rekrutierung und liessen sie in allen deutschschweizer Rekrutierungszentren während zwei Wochen anstelle der Forced-Choice-Variante von den Stellungspflichtigen bearbeiten. Mit dieser Datenerhebung verfolge ich zudem zwei weitere Ziele: Ich will aufzeigen, dass sich die Reliabilität der drei Skalen mit diesem Antwortformat deutlich verbessern lässt und es dadurch möglich ist, die Anzahl der Items von zehn auf acht pro Skala zu reduzieren. In Tabelle 7.7 habe ich die Zusammensetzung der Stichprobe der Testdurchführung mit der likert-skalierten Version des Leadership-Fragebogens aufgeführt. In Klammern steht jeweils die Anzahl der auf Grund unten beschriebener Kriterien ausgeschlossenen Datensätze.

Tabelle 7.7
Stichprobe der Überprüfungsstudie mit der likert-skalierten Version des Leadership-Fragebogens

Rekrutierungs-zentrum	Woche vom 20.8.07		Woche vom 27.8.07		Total		%
Sumiswald	151	(11)	166	(9)	317	(20)	31.17
Nottwil	82	(6)	70	(12)	152	(18)	14.95
Windisch	164	(10)	117	(9)	281	(19)	27.63
Rüti	66	(13)	95	(11)	161	(24)	15.83
Mels	17	(0)	89	(7)	106	(7)	10.42
Total					1'017	(88)	100.00

Anmerkung. In Klammern ist jeweils die Anzahl der ausgeschlossenen Datensätze aufgeführt.

Für die nachfolgenden Berechnungen führe ich beim Ausgangsdatensatz (N = 1'105) dieselbe Datenbereinigung durch, wie unter Kapitel 7.1 für den Datensatz 2008 beschrieben. Wiederum wende ich folgende drei Kriterien an:

1. Testbearbeitungszeit: Anhand von Box-Plots eruiere ich Stellungspflichtige mit einer sehr kurzen (< 354 sec; n = 13, 1.2%) oder sehr langen (> 1'760 sec; n = 16, 1.5%) Testbearbeitungszeit und schliesse diese aus dem Datensatz aus.
2. Durchklicken, Variante 1: Insgesamt 55 (5.1%) Datensätze schliesse ich auf Grund einer sehr häufig oder sehr selten gewählten Antwortalternative aus. Folgende Ausschlusskriterien gelangen zum Einsatz:
   Antwortalternative 1: > 46 Mal gewählt (n = 22)
   Antwortalternative 2: > 64 oder < 5 Mal gewählt (n = 14; n = 3)
   Antwortalternative 3: > 73 oder < 9 Mal gewählt (n = 11; n = 5)
   Antwortalternative 4: > 59 Mal gewählt (n = 7)

3. Durchklicken, Variante 2: Anhand der Berechnung der Streuung der Antworten pro Dimension eruiere ich diejenigen Stellungspflichtigen, welche den Fragebogen unseriös bearbeitet haben. Nur vier (0.4%) von ihnen schliesse ich auf Grund einer zu grossen Streuung (*SD* > 1.33) in der Skala Kontaktfähigkeit aus.

Somit schliesse ich die Daten von insgesamt 88 (8.0%) Stellungspflichtigen aus, was zu einem Datensatz mit 1'017 Probanden führt. Zuerst gehe ich anhand dieses Datensatzes der Frage nach, ob sich die drei Skalen auch mit dem likert-skalierten Itemformat faktorenanalytisch abbilden lassen. Der in Abbildung 7.5 (links) dargestellte Scree-Plot weist jedoch auf fünf Faktoren hin, welche nach der Varimax-Methode rotiert nur gerade 27.99% der Varianz aufklären.

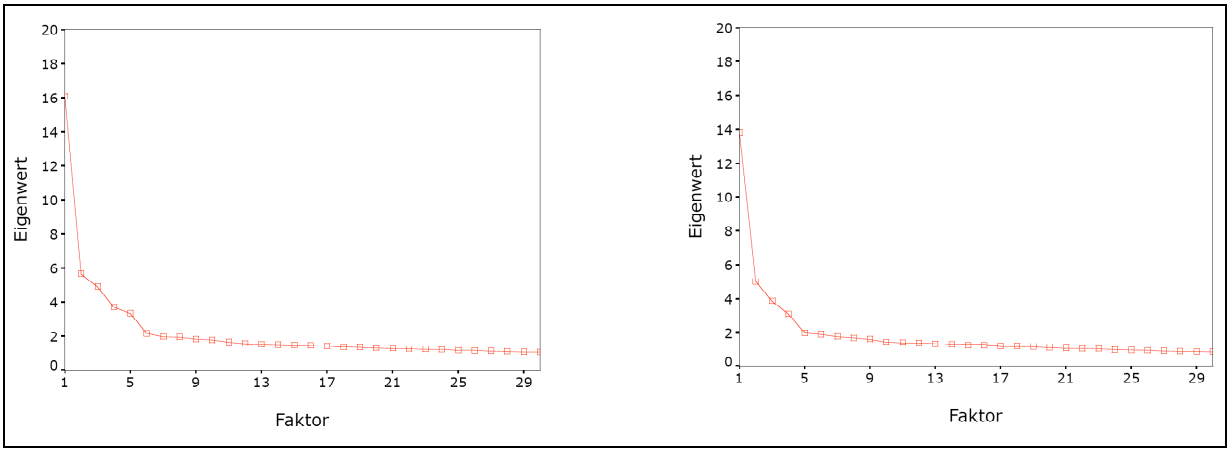


Abbildung 7.5 Scree-Plots der Faktorenanalyse der likert-skalierten Version des Leadership-Fragebogens (vier resp. drei Wertequadranten; 120 resp. 90 Items).

Anhand der in der rotierten Komponentenmatrix enthaltenen Ladungen der einzelnen Items auf die fünf Faktoren suche ich nach einer inhaltlichen Erklärung für die fünf Faktoren. In einer vereinfachten Form habe ich die Komponentenmatrix in Tabelle 7.8 dargestellt. Es lässt sich erkennen, dass die drei Leadership-Skalen in unterschiedlichem Ausmass auf die Faktoren 1 (90% der Kontaktfähigkeits-Items), 3 (55% der Verantwortungsbewusstseins-Items) und 5 (58% der Durchsetzungsfähigkeits-Items) laden. Interessant sind die Faktoren 2 und 4: Auf Faktor 2 laden 60% der Items des Wertequadranten 2 (alternative Verhaltensweise); von der Skala Durchsetzungsfähigkeit und Verantwortungsbewusstsein sind es sogar 80%. Auf den Faktor 4 laden neun Items der Wertequadranten 3 und 4 der Skala Durchsetzungsfähigkeit negativ und fünf Items der Wertequadranten 1 und 2 der Skala Verantwortungsbewusstsein positiv.

Tabelle 7.8

Vereinfachte Darstellung der rotierten Komponentenmatrix der Fünf- und der Drei-Faktoren-Lösung der Items des likert-skalierten Leadership-Fragebogens

		Fünf-Faktoren-Lösung				Drei-Faktoren-Lösung			
	Wertequadranten	1	2	3	4	1	2	3	4
Durchsetzungsfähigkeit	Lohnerhöhung	5	2	5	4	3	2	3	3
	Fahrer	5	5	5	5	3	3	3	3
	Unterbruch	5	2	4	4	3	2	3	3
	Aufräumen	5	2	5	5	2	2	3	3
	Waschküche	5	2	4	4	3	2	3	3
	Geschirr	5	2	2	5	2	2	2	3
	Musik	5	2	5	4	3	2	3	2
	Problem	5	2	5	4	3	2	3	3
	Auswärts	5	5	4	4	3	2	3	3
	Arbeit	5	5	5	5	3	3	3	3
Kontaktfähigkeit	Nachbarn	1	1	1	1	2	1	2	1
	Zugfahrt	1	1	1	1	2	1	1	1
	Kurs	1	1	1	1	1	1	1	1
	Schultag	1	2	2	1	1	2	2	1
	Flugzeug	1	2	1	1	1	2	1	1
	Barmann	1	1	1	1	1	1	1	1
	Begleitung	1	1	1	1	1	1	1	1
	Umzug	1	1	2	1	2	1	2	1
	Geburtstag	1	1	1	1	1	1	1	1
	Fitness	1	1	1	1	1	1	1	1
Verantwortungs-bewusstsein	Schanze	3	2	3	2	2	2	2	2
	Beratungsstelle	3	2	3	3	2	2	2	1
	Silvester	3	2	3	3	2	2	2	2
	Subvention	4	2	3	3	2	2	2	1
	Nachhilfestunden	4	2	2	3	2	2	2	2
	Kind	4	2	3	3	3	2	2	1
	Wohnung	3	2	2	3	2	2	2	3
	Bergtour	4	2	3	3	2	2	2	2
	Ampel	3	4	3	3	2	2	2	1
	Autofahren	4	2	3	3	2	2	2	2

Anmerkung. Die Zahlen in den Ladungs-Feldern stehen für die extrahierten Faktoren.

Ich erkläre mir die Faktoren 2 und 4 durch den Einsatz der Wertequadrate bei der Item-Konstruktion. Der Faktor 4 könnte jedoch auch Selbstdarstellungstendenzen abbilden: Wer in der Armee keine Kaderposition übernehmen möchte, der lehnt stark durchsetzungsfähiges Verhalten ab und stellt sich als total verantwortungslos dar. Anhand einer forcierten Drei-Faktoren-Lösung untersuche ich, ob sich die inhaltlich – mit den drei Skalen des Fragebogens – oder die zwei methodisch erklärbaren Faktoren durchsetzen. Wie zu erwarten ist, zeigen sich die drei Skalen nun deutlicher. Die drei Faktoren klären wiederum mit 22.15% nur wenig Varianz auf. Auffällig – und in Tabelle 7.8 gut ersichtlich – ist, dass zwölf Items der Skala Durchsetzungsfähigkeit und acht Items der Skala Kontaktfähigkeit auf den Faktor der Skala Verantwortungsbewusstsein laden, wovon die

Hälfte dieser Items zum zweiten Wertequadranten gehört. Somit lässt sich auch für diese Lösung keine befriedigende und vollständige Erklärung finden.

Bei beiden bisher gerechneten Varianten scheint der zweite Wertequadrant problematisch zu sein. Ich rechne aus diesem Grund nochmals eine Faktorenanalyse, bei welcher ich diesen ausschliesse. Der Scree-Plot (Abbildung 7.5 rechts) weist dann nur noch auf vier Faktoren hin, welche zusammen 28.65% der Varianz erklären. In Tabelle 7.9 stelle ich die entsprechend angepasste Komponentenmatrix dar, welche wiederum den die beiden Skalen Durchsetzungsfähigkeit und Verantwortungsbewusstsein verbindenden Faktor enthält. Die in derselben Tabelle dargestellte forcierte Drei-Faktoren-Lösung bildet nun schon ziemlich gut die drei Faktoren des Fragebogens ab.

Tabelle 7.9

Vereinfachte Darstellung der rotierten Komponentenmatrix der Fünf- und der Drei-Faktoren-Lösung der Items des likert-skalierten Leadership-Fragebogens

		Vier-Faktoren-Lösung				Drei-Faktoren-Lösung			
	Wertequadranten	1	2	3	4	1	2	3	4
Durchsetzungsfähigkeit	Lohnerhöhung	4		4	3	2		3	3
	Fahrer	4		4	4	3		3	3
	Unterbruch	4		3	4	3		3	3
	Aufräumen	4		4	4	2		3	3
	Waschküche	4		3	3	3		3	3
	Geschirr	4		3	3	2		3	3
	Musik	4		4	3	3		3	2
	Problem	2		4	3	3		3	3
	Auswärts	4		3	3	3		3	3
	Arbeit	4		4	3	3		3	3
Kontaktfähigkeit	Nachbarn	1		1	1	1		1	1
	Zugfahrt	1		1	1	1		1	1
	Kurs	1		1	1	1		1	1
	Schultag	1		1	1	1		1	1
	Flugzeug	1		1	1	1		1	1
	Barmann	1		1	1	1		1	1
	Begleitung	1		1	1	1		1	1
	Umzug	1		1	1	2		1	1
	Geburtstag	1		1	1	1		1	1
	Fitness	1		1	1	1		1	1
Verantwortungsbewusstsein	Schanze	2		2	2	2		2	2
	Beratungsstelle	2		2	2	2		2	1
	Silvester	2		2	2	2		1	2
	Subvention	3		2	2	2		1	1
	Nachhilfestunden	3		2	2	2		2	1
	Kind	3		2	2	2		1	1
	Wohnung	2		2	2	2		2	2
	Bergtour	3		2	2	2		2	2
	Ampel	2		2	2	2		2	1
	Autofahren	2		2	2	2		2	2

Anmerkung. Die Zahlen in den Ladungs-Feldern stehen für die extrahierten Faktoren.

In der nachfolgend aufgeführten Tabelle 7.10 stelle ich die Kennwerte für die Einschätzung der Eignung der Daten für die Durchführung einer Faktorenanalyse dar. Der Wert nach Kaiser-Meyer-Olkin zeigt an, dass die Variablen genügend hoch miteinander korrelieren und die Korrelationskoeffizienten gemäss dem Bartlett-Test auf Sphärizität signifikant von Null verschieden sind. Es scheint sogar so zu sein, dass die Variablen zu hoch miteinander korrelieren und Multikollinearität vorliegen könnte, wie sich anhand der Determinante der Korrelationsmatrix und dem Test nach Haitovsky (1969) ableiten lässt. Hohe Korrelationen zwischen den Items können dazu führen, dass sich diese nicht eindeutig einem Faktor zuordnen lassen. Gemäss Field (2009) ist jedoch Multikollinearität bei einer Hauptkomponentenanalyse vernachlässigbar. Interessanterweise zeigt die Analyse der Item-Korrelationsmatrix, dass kein Korrelationskoeffizient über .70 liegt – im Gegenteil: Die überwiegende Anzahl liegt unter .30, was ebenfalls auf eine ungünstige Ausgangslage hindeutet. Zusammenfassend lässt sich feststellen, dass sich der vorliegende Datensatz nicht genügend gut für die Durchführung einer Faktorenanalyse eignet und die oben präsentierten Resultate aus diesem Grund mit der notwendigen Vorsicht zu interpretieren sind.

Tabelle 7.10

Kennwerte der Eignung der Daten für eine Faktorenanalyse und Angaben zu den Varimax-rotierten Lösungen

	4 Wertequadranten pro Item (Tabelle 7.8)	3 Wertequadranten pro Item (Tabelle 7.9)
KMO	.90	.90
Kleinster KMO-Wert (Grenzwert .50)	.62	.69
Bartlett–Test	χ^2 = 40'852, *df* = 7'140, *p* < .001	χ^2 = 29'838, *df* = 4'005, *p* < .001
Determinante der Korrelationsmatrix (Grenzwert .00001)	|*R*| < 0.00001	|*R*| < 0.00001
Test nach Haitovsky	χ^2_H < 0.01, *df* = 7'140, *p* > .05	χ^2_H < 0.01, *df* = 4'005, *p* > .05
Anzahl Iterationen	15 (6)	11 (6)
Erklärte Varianz	27.99% (25.22%)	28.65% (25.20%)

Anmerkung. Die Zahlen in Klammern stehen jeweils für Durchführung der um zwei Faktoren (Tabelle 7.8) respektive einen Faktor (Tabelle 7.9) reduzierten Analyse.

Auf Grund der potenziell ungünstigen Ausgangslage der Datenmatrix für die Durchführung einer Faktorenanalyse und der doch uneindeutigen Faktorenstruktur der likert-skalierten Version des Leadership-Fragebogens führe ich die nachfolgende Reduktion der Anzahl Items für jede Skala einzeln anhand der Analyse der Trennschärfen der Items durch. Dazu musste ich in einem ersten Schritt die Items des Wertequadranten 1 und vereinzelte des Wertequadranten 2 umpolen,

um so unipolare Skalen erzeugen zu können. Die von dieser Prozedur betroffenen Sub-Items habe ich empirisch anhand der jeweiligen Trennschärfen bestimmt.

Tabelle 7.11 zeigt die Reliabilitäten der gekürzten Versionen der Skalen des likert-skalierten Leadership-Fragebogens. Bei der Kürzung bin ich iterativ vorgegangen, indem ich jeweils das Item mit der tiefsten durchschnittlichen – über alle vier respektive drei Verhaltensweisen gemittelte – Trennschärfe ausgeschlossen und dann erneut die Trennschärfen und Reliabilitäten berechnet habe. Dieses Vorgehen habe ich auf Grund oben dargestellter Analysen mit einer Testversion mit allen vier Wertequadranten und einer mit ausgeschlossenem zweiten Wertequadrant getrennt durchgeführt. In den Anhängen 7.7 bis 7.9 habe ich die jeweiligen Itemkennwerte detailliert abgedruckt.

Tabelle 7.11
Reliabilitäten (Cronbach Alpha) der gekürzten Versionen des likert-skalierten Leadership-Fragebogens

	F-C-Variante	4 Wertequadranten pro Item					3 Wertequadranten pro Item				
Anzahl Item-Stämme	10	10	8	6	*10*	*6*	10	8	6	*10*	*6*
Anzahl Sub-Items	10	40	32	24	*(10)*	*(6)*	30	24	18	*(10)*	*(6)*
Durchsetzungsfähigkeit	.57	.81	.80	.78	*.72*	*.65*	.81	.79	.77	*.73*	*.66*
Kontaktfähigkeit	.84	.93	.92	.91	*.91*	*.88*	.93	.92	.91	*.91*	*.88*
Verantwortungsbewusstsein	.70	.86	.85	.83	*.79*	*.73*	.84	.83	.81	*.78*	*.73*

Anmerkung. *N* = 1'017. Die unter der Rubrik F-C-Variante aufgeführten Werte beziehen sich auf die anhand des Datensatzes 2008 (*N* = 19'801) mit der Forced-Choice-Variante des Fragebogens berechneten Reliabilitäten. In kursiv sind die Werte gesetzt, welche sich ergeben, wenn die vier Sub-Items (Wertequadranten) pro Item zusammengefasst in die Reliabilitäts-Berechnung einfliessen.

Die in Tabelle 7.11 dargestellten Werte belegen eindrücklich – und wie nicht anders erwartet – die drastische Erhöhung der Reliabilitäten durch die Verwendung einer Likert-Skalierung im Vergleich zur Forced-Choice-Variante: So liegen jetzt alle Werte über der Grenze von α = .80, was gemäss den Angaben in Tabelle 6.16 als gut bezeichnet werden darf. Auch die Verkürzung der Skalen auf acht respektive sechs Items führt noch zu akzeptablen Ergebnissen und einzig der Wert der Skala Durchsetzungsfähigkeit fällt knapp unter die angestrebte Grenze. Dabei macht es keinen Unterschied, ob sich die Items jeweils aus vier oder drei Wertequadranten zusammensetzen. Hingegen sinken die Reliabilitätswerte wieder, wenn ich pro Item nur ein aus den Einstufungen zu den vier respektive drei Wertequadranten gebildeten Summenwert in die Berechnungen einfliessen

lasse. Der Vergleich mit den Reliabilitäten der Forced-Choice-Skalen zeigt aber auch in diesem Fall eine deutliche Verbesserung. Die Abkehr von der bei Situational Judgment Tests üblichen Wahl der am ehesten dem eigenen Verhalten entsprechenden Verhaltensalternative und die Einführung einer Einstufung jeder Verhaltensalternative auf einer Likert-Skala führt somit in jedem Fall zu einer Verbesserung der Reliabilitäten der Leadership-Skalen. Da die Psychologen der Rekrutierungszentren der Schweizer Armee den Test nur als grobes Screening-Instrument im Rahmen einer Kader-Vorselektion einsetzen, wäre es aus meiner Sicht zulässig, den Test von insgesamt 30 Items auf 18 Items zu reduzieren. Tabelle 7.12 lässt sich entnehmen, dass sich damit die durchschnittliche Testbearbeitungszeit von derzeit knapp 15 Minuten auf ca. 10 Minuten verkürzen liesse. In dieser Tabelle wird auch ersichtlich, dass – wie auch von Testassistenten in den Rekrutierungszentren berichtet – das Bearbeiten der likert-skalierten Version mehr Zeit in Anspruch nimmt, nämlich knapp 25% mehr, als das beim Forced-Choice-Verfahren der Fall ist.

Tabelle 7.12
Testbearbeitungszeiten (Datensätze 2008 Forced-Choice und 2007 Likert)

	N	min	max	*M*	*SD*
Forced-Choice, gesamt (3x30 Items)	21'008	0.10	104.05	14.59	4.04
Forced-Choice, von Ausreissern bereinigt	19'801	4.42	28.57	14.52	3.49
Likert-skaliert, gesamt (3x40 Verhaltensweisen)	1'105	1.09	64.59	18.00	4.88
Likert-skaliert, von Ausreissern bereinigt	1'017	6.80	29.17	18.00	4.06
Likert-skaliert (3x32 Verhaltensweisen)			*23.34*	*14.40*	
Likert-skaliert (3x24 Verhaltensweisen)			*17.50*	*10.80*	
Likert-skaliert (3x18 Verhaltensweisen)			*13.13*	*8.10*	

Anmerkung. Bei den in kursiv gesetzten Zeiten handelt es sich um berechnete Schätzwerte.

Mit dem auf sechs Items pro Skala reduzierten Leadership-Fragebogen habe ich nochmals Faktoren-Analysen gerechnet. Die in Abbildung 7.6 dargestellten Scree-Plots zu den Varianten mit vier respektive drei Wertequadranten weisen wiederum – jedoch weniger deutlich als bei den oben geschilderten Analysen – auf eine Fünf- respektive Vier-Faktoren-Lösung hin. In Tabelle 7.13 habe ich die forcierten Drei-Faktoren-Lösungen dargestellt. Es ist ersichtlich, dass die drei Skalen des Leadership-Fragebogens schon relativ gut abgebildet werden, die faktorielle Validität jedoch mit ca. 20% respektive 15% Items mit der höchsten Ladung auf einem fremden Faktor noch keineswegs zufrieden stellend ist. Zudem zeigt sich das Problem der Multikollinearität auch hier wieder (siehe Tabelle 7.14).

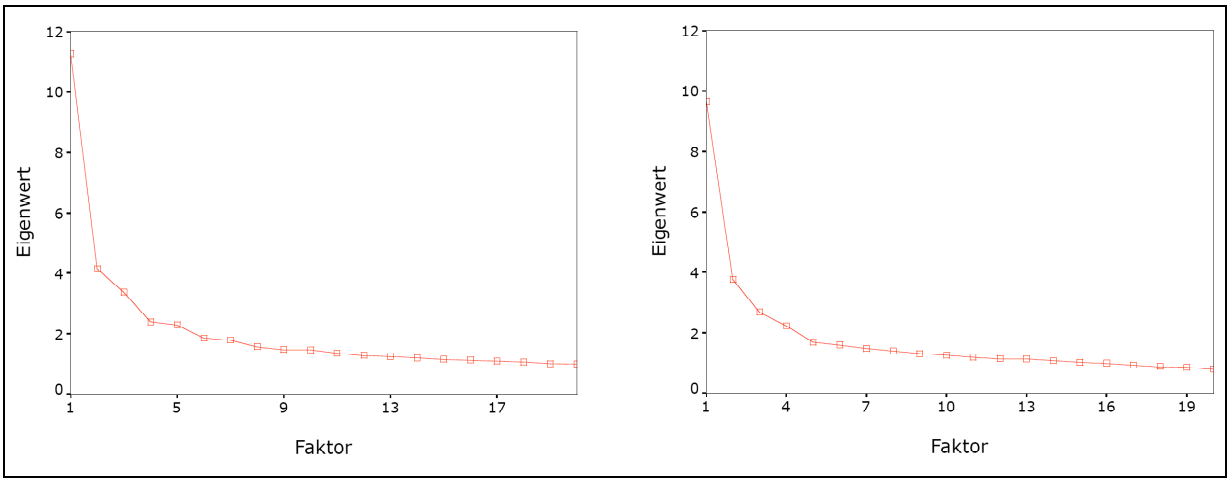


Abbildung 7.6 Scree-Plots der Faktorenanalysen der Vier- und Drei-Wertequadranten-Versionen des gekürzten likert-skalierten Leadership-Fragebogens.

Tabelle 7.13

Vereinfachte Darstellung der rotierten Komponentenmatrix der Drei-Faktoren-Lösungen der Versionen mit vier respektive drei Wertequadranten des gekürzten likert-skalierten Leadership-Fragebogens

		4 Wertequadranten				3 Wertequadranten			
	Wertequadranten	1	2	3	4	1	2	3	4
Durchsetzungs-fähigkeit	Fahrer	3	3	3	3	3		3	3
	Unterbruch	3	1	3	3	3		3	3
	Aufräumen	1	1	3	3	2		3	3
	Waschküche	3	1	3	3	3		3	3
	Geschirr					2		3	3
	Auswärts	3	1	3	3				
	Arbeit	3	3	3	3	3		3	3
Kontakt-fähigkeit	Nachbarn	1	2	1	2				
	Zugfahrt	1	2	2	2	1		1	1
	Flugzeug					1		1	1
	Begleitung	2	2	2	2	1		1	1
	Umzug	1	2	1	2	2		1	1
	Geburtstag	1	2	2	2	1		1	1
	Fitness	2	2	2	2	1		1	1
Verantwortungsbewusstsein	Beratungsstelle	1	1	1	2	2		2	1
	Silvester	1	1	1	1	2		2	2
	Subvention	1	1	1	2	2		1	1
	Nachhilfestunden	1	1	1	1	2		2	1
	Ampel	1	1	1	2	2		2	1
	Autofahren	1	1	1	1	2		2	2

Anmerkung. Die Zahlen in den Ladungs-Feldern stehen für die extrahierten Faktoren. Grau unterlegt sind Ladungsfelder von Items, welche in der entsprechenden Version nicht enthalten sind.

Tabelle 7.14

Kennwerte der Eignung der Daten für eine Faktorenanalyse und Angaben zu den Varimax-rotierten Lösungen

	4 Wertequadranten	3 Wertequadranten
KMO	.89	.89
Kleinster KMO-Wert (Grenzwert .50)	.70	.69
Bartlett–Test	χ^2 = 23'095, *df* = 2'556, *p* < .001	χ^2 = 17'420, *df* = 1'431, *p* < .001
Determinante der Korrelationsmatrix (Grenzwert .00001)	|*R*| < 0.00001	|*R*| < 0.00001
Test nach Haitovsky	χ^2_H < 0.01, *df* = 2'556, *p* > .05	χ^2_H < 0.01, *df* = 1'431, *p* > .05
Anzahl Iterationen	5	5
Erklärte Varianz	26.06%	29.82%

Da die oben berichtete Faktorenanalyse nicht zu befriedigenden Resultaten führt, rechne ich abschliessend noch eine über die 18 Items. Dazu habe ich zuerst die Einstufungen der vier respektive drei Wertequadranten zu einem Wert pro Item aufsummiert. Ich erhoffe mir damit, dass ich so die intervenierenden Methoden- und/oder Antwortverhaltenseinflüsse ausschalten kann. Die Scree-Plots (Abbildung 7.7) der beiden Versionen (vier respektive drei Wertequadranten pro Item) weisen dann auch klar auf zwei Drei-Faktoren-Lösungen hin, welche zusammen je knapp 50% Varianz erklären. In der in Tabelle 7.15 dargestellten Matrizen mit den Faktorladungen zeigt sich, dass jedes Item auf den richtigen Faktor lädt und dass lediglich zwei respektive ein Item noch eine Nebenladung auf einen zweiten Faktor aufweisen.

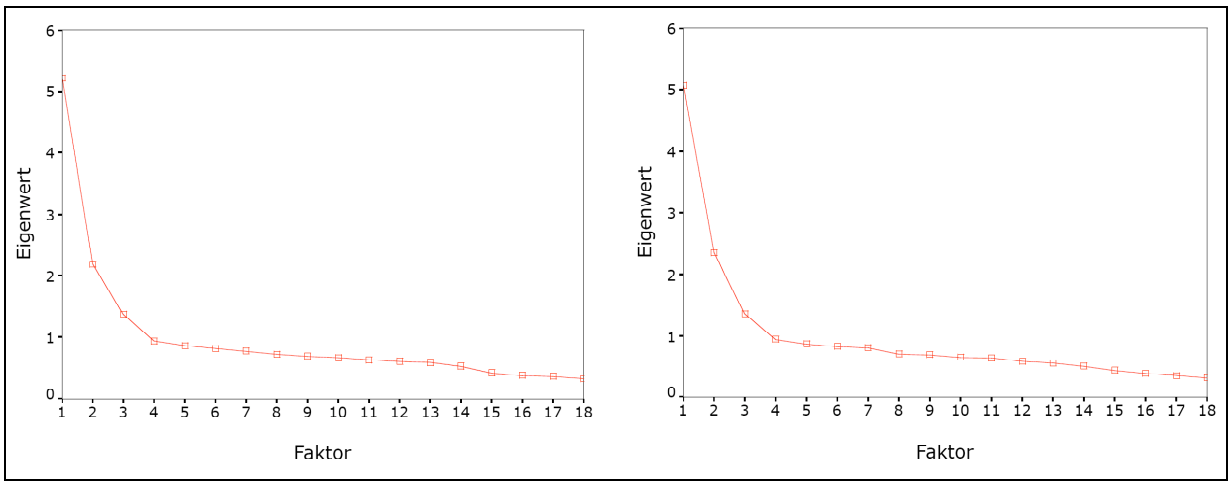


Abbildung 7.7 Scree-Plots der Faktorenanalysen der Vier- und Drei-Wertequadranten-Versionen des gekürzten Leadership-Fragebogens.

Tabelle 7.15

Matrizen der Faktorladungen der Drei-Faktoren-Lösungen der Versionen mit vier respektive drei Wertequadranten des gekürzten likert-skalierten Leadership-Fragebogens

4 Wertequadranten pro Item				3 Wertequadranten pro Item			
	Faktor 1 KF	Faktor 2 VB	Faktor 3 DF		Faktor 1 KF	Faktor 2 VB	Faktor 3 DF
Zugfahrt	**.82**	.14	.01	Flugzeug	**.81**	.19	.03
Schultag	**.81**	.20	-.02	Schultag	**.81**	.20	.04
Flugzeug	**.81**	.18	-.03	Zugfahrt	**.79**	.16	.06
Begleitung	**.77**	.25	-.09	Begleitung	**.76**	.24	-.08
Barmann	**.64**	**.45**	-.06	Fitness	**.72**	.27	-.06
Kurs	**.62**	**.36**	-.01	Kurs	**.62**	**.33**	.01
Subvention	.09	**.70**	.04	Subvention	.12	**.68**	.07
Kind	.16	**.68**	.03	Kind	.15	**.66**	.04
Wohnung	.19	**.64**	-.07	Wohnung	.18	**.64**	-.07
Silvester	.24	**.62**	-.04	Silvester	.26	**.61**	.00
Ampel	.20	**.62**	.03	Ampel	.24	**.58**	.04
Bergtour	.27	**.48**	-.10	Bergtour	.25	**.47**	-.08
Auswärts	.02	.04	**.64**	Auswärts	.06	.20	**.72**
Waschküche	-.09	-.09	**.63**	Lohnerhöhung	.09	.29	**.68**
Musik	-.03	-.10	**.62**	Problem	-.03	-.04	**.59**
Arbeit	.04	.05	**.61**	Arbeit	-.08	-.05	**.59**
Problem	.05	.11	**.56**	Waschküche	-.02	-.14	**.57**
Aufräumen	-.17	-.12	**.55**	Musik	.02	-.06	**.56**
Eigenwert	3.65	2.89	2.22		3.69	2.74	2.33
erklärte Varianz	20.30	16.08	12.34		20.53	15.24	12.96

Anmerkung. DF = Durchsetzungsfähigkeit; KF = Kontaktfähigkeit; VB = Verantwortungsbewusstsein.

Tabelle 7.16

Kennwerte der Eignung der Daten für eine Faktorenanalyse und Angaben zu den Varimax-rotierten Lösungen

	4 Wertequadranten pro Item	3 Wertequadranten pro Item
KMO	.90	.89
Kleinster KMO-Wert (Grenzwert .50)	.72	.73
Bartlett–Test	χ^2 = 5'271, df = 153, p < .001	χ^2 = 5'314, df = 153, p < .001
Determinante der Korrelationsmatrix (Grenzwert .00001)	$\|R\|$ = 0.00539	$\|R\|$ = 0.00517
Test nach Haitovsky	χ^2_H = 5.45, df = 153, p > .05	χ^2_H = 5.23, df = 153, p > .05
Anzahl Iterationen	5	5
Erklärte Varianz	48.71%	48.72%

Auf die Korrelationen zwischen den drei Skalen wirkt sich das Skalenformat praktisch nicht aus, wie Tabelle 7.17 zu entnehmen ist. Dass sich die Korrelationen zwischen der Skala Durchsetzungsfähigkeit und den beiden anderen Skalen in der Version mit den drei Wertequadranten leicht ändern, diejenige zwischen der Skala Kontaktfähigkeit und der Skala Verantwortungsbewusstsein jedoch gleich bleibt, lässt sich als Anzeichen dafür deuten, dass sich die Skala Durchsetzungsfähigkeit inhaltlich minimal verändert, wenn der zweite Wertequadrant unberücksichtigt bleibt. Die durchschnittlichen Interkorrelationen zwischen den Items in den drei Auswertungen unterscheiden sich praktisch kaum und fallen bei der Skala Kontaktfähigkeit eher hoch aus, was auf eine tendenziell eng gefasste Operationalisierung des Konstruktes hindeutet (Buss & Craik, 1983).

Tabelle 7.17
Vergleich der Datensätze 2008 und 2007: Korrelationsmatrizen

	Datensatz 2008; Forced-Choice; 10 Items pro Skala (*N* = 19'801)			Datensatz 2007; likert-skaliert; 6 Items pro Skala (*N* = 1'017)					
				4 Wertequadranten			3 Wertequadranten		
Skala	DF	KF	VB	DF	KF	VB	DF	KF	VB
Durchsetzungsfähigkeit	*.12*			*.12*			*.16*		
Kontaktfähigkeit	-.08	*.34*		-.11	*.29*		.01	*.36*	
Verantwortungsbewusstsein	-.03	.53	*.20*	-.09	.58	*.17*	.05	.57	*.20*

Anmerkung. Alle Korrelationen des Datensatzes 2008 sind hochsignifikant. Im Datensatz 2007 sind Korrelationen >.06 signifikant, > .10 sehr signifikant und > .50 hochsignifikant (Bonferroni-korrigiert). Kursiv in der Diagonale sind die durchschnittlichen Itemkorrelationen aufgeführt. DF = Durchsetzungsfähigkeit; KF = Kontaktfähigkeit; VB = Verantwortungsbewusstsein.

Anhand eines für die Erhebung des Bekanntheitsgrades der in den Items geschilderten Situationen und zur Bestimmung des Zusammenhangs der Leadership-Dimensionen mit den Big Five-Persönlichkeitsfaktoren erhobenen Datensatzes kann ich die Korrelationen zwischen den beiden Fragebogen-Versionen bestimmen. Dafür stehen mir die Angaben von 100 Studierenden der Psychologie zur Verfügung, die jeweils eines der beiden ausbalancierten Fragebogenhefte – bestehend aus den beiden Leadership-Versionen und dem NEO-PI-R – bearbeitet haben[1]. In Tabelle 7.18 habe ich die Interkorrelationen zwischen den drei Leadership-Dimensionen getrennt nach der Forced-Choice- und der likert-skalierten Version und die entsprechenden Reliabilitätskoeffizienten aufgeführt. Wie dort ersichtlich ist, fallen letztere ähnlich aus, wie diejenige, welche ich anhand

[1] Eine ausführliche Beschreibung der Stichprobe findet sich in Kapitel 7.3. In Anhang 7.11 ist ein likert-skaliertes Item des dabei eingesetzten Fragebogens abgebildet.

der Datensätze 2007 und 2008 aus den Rekrutierungszentren berechnet habe. Nur bei der Forced-Choice-Version der Dimension Verantwortungsbewusstsein zeigt sich ein grosser Unterschied: Bei den Stellungspflichtigen errechnete ich ein Cronbach Alpha von .70, bei den Studierenden eines von .49. Auch bei den Korrelationen zwischen den drei Skalen ergeben sich Unterschiede bei der Dimension Verantwortungsbewusstsein – der grösste wiederum in der Forced-Choice-Version bei der Korrelation zur Kontaktfähigkeit, welche bei den Stellungspflichtigen r = .69, bei den Studierenden jedoch nur r = .27 beträgt. Auch der Vergleich dieser Korrelation zwischen den beiden Fragebogen-Versionen bestätigt, dass die Studierenden die Items der Skala Verantwortungsbewusstsein anders eingeschätzt haben als erwartet: Die Korrelation zwischen dieser Skala und der Skala Kontaktfähigkeit beträgt bei den Studierenden in der Forced-Choice-Version (gelb unterlegt) r = .27, in der likert-skalierten Version (grün unterlegt) r = .47. Bei den Stellungspflichtigen unterscheiden sich diese beiden Werte nur geringfügig (r = .69 [rot] vs. r = .65 [blau]). Somit lässt sich aussagen, dass die bei den Studierenden erhobenen Daten mit denjenigen der Stellungspflichtigen vergleichbar sind mit Ausnahme derjenigen der Skala Verantwortungsbewusstsein der Forced-Choice-Version.

Tabelle 7.18

Korrelationsmatrix der beiden Leadership-Fragebogen-Versionen (zehn Items pro Skala; Vergleich der Datensätze aus den Rekrutierungszentren mit dem der Psychologie-Studierenden)

	Forced-Choice			likert-skaliert		
	DF	KF	VB	DF	KF	VB
Durchsetzungs-fähigkeit	*.57 / .62*	-.07 (-.10)	.17 (.31)	*.81 / .84*	-.04 (-.05)	.12 (.15)
Kontaktfähigkeit	-.08* (-.12)	*.84 / .79*	.17 (.27)	-.06 (-.07)	*.93 / .93*	.41* (.47)
Verantwortungs-bewusstsein	-.03* (-.05)	.53* (.69)	*.70 / .49*	.00 (.00)	.58* (.65)	*.86 / .81*

Anmerkung. Die rot und blau unterlegten Korrelationen beziehen sich auf die Datensätze 2008 (Forced-Choice; N = 19'801) resp. 2007 (likert-skaliert; N = 1'017) aus den Rekrutierungszentren, die gelb und grün hinterlegten auf den Studierenden-Datensatz (N = 100). In Klammern sind die doppelt minderungskorrigierten Korrelationen angegeben. In der Diagonale stehen die Reliabilitäten (Cronbach Alpha) der Skalen. Bonferroni-korrigiertes Signifikanzniveau p < .02. DF = Durchsetzungsfähigkeit; KF = Kontaktfähigkeit; VB = Verantwortungsbewusstsein.

Die Verteilungskennwerte der Skalen des Leadership-Fragebogens in den Datensätzen der Rekrutierungszentren und der Psychologie-Studierenden habe ich in Anhang 7.10 aufgeführt. Es zeigt sich, dass die Studierenden tiefere Durch-

schnittswerte erzielen als die Stellungspflichtigen. Zudem stellen sie sich in der Forced-Choice-Version gemässigter dar, was sich auch in den tieferen Streuungen der Werte zeigt. Beide Beobachtungen lassen sich grundsätzlich auf das Antwortverhalten der Stellungspflichtigen in der Selektionssituation zurückführen, wobei hier Verzerrungen in beide Richtungen auftreten, da es sowohl Stellungspflichtige gibt, welche auf keinen Fall eine militärische Kaderlaufbahn einschlagen wollen, als auch solche, welche sich die Übernahme einer Kaderposition wünschen. Die Abweichungen zwischen den beiden Stichproben lassen sich bis zu einem gewissen Grad auch durch die Stichprobengrösse erklären: Da Extremwerte in den drei Skalen eher selten vorkommen, ist deren Auftretenswahrscheinlichkeit in einer kleinen Stichprobe tief. Dies habe ich in Anhang 7.10 entsprechend simuliert, indem ich die Skalenkennwerte anhand 100 zufällig aus dem 19'801 Stellungspflichtige umfassenden Datensatz ausgewählter Probanden berechnet habe. Auch in dieser Simulation zeigen sich die Unterschiede zur Studentenpopulation, jedoch treten nun auch hier kaum noch Extremwerte auf.

Die anhand des Datensatzes der Psychologie-Studierenden berechneten Korrelationen zwischen den zwei Leadership-Fragebogen-Versionen habe ich in Tabelle 7.19 aufgeführt. Interessant ist hierbei, dass sich die Korrelationen zwischen den beiden Versionen in Abhängigkeit der jeweiligen Skala deutlich unterscheiden: So betragen die gerundeten Korrelationskoeffizienten bei der Skala Kontaktfähigkeit r = .90, bei der Skala Durchsetzungsfähigkeit r = .80 und bei der Skala Verantwortungsbewusstsein r = .60. Die mit der Formel von Spearman (siehe z. B. Lienert & Raatz, 1998) durchgeführte doppelte Minderungskorrektur (von Spearman als *Attenuationskorrektur* bezeichnet) ergibt für alle drei Skalen eine perfekte Übereinstimmung der beiden Versionen, womit ich belegen kann, dass hauptsächlich die ungenügende Reliabilität der Forced-Choice-Version der Skala Verantwortungsbewusstsein für die tiefen Korrelationen verantwortlich ist.

Tabelle 7.19
Korrelationsmatrix der beiden Leadership-Versionen

		likert-skaliert		
		Durchsetzungs-fähigkeit	Kontaktfähigkeit	Verantwortungs-bewusstsein
Forced-Choice	Durchsetzungsfähigkeit	**.79 (1.00)*** **	-.06 (-.08)	-.04 (-.06)
	Kontaktfähigkeit	-.07 (-.09)	**.88 (1.00)*** **	.34 (.43)***
	Verantwortungsbewusstsein	.20 (.31)	.25 (.37)	**.64 (1.00)*** **

Anmerkung. N = 100. In Klammern sind die doppelt minderungskorrigierten Korrelationen angegeben. Bonferroni-korrigiertes Signifikanzniveau *** p < .006.

Der weiter oben dargestellte Vergleich der Resultate der Faktorenanalysen mit dem likert-skalierten und dem Forced-Choice-Format zeigt auf, dass der Einbezug der Antworten auf dem Niveau der alternativen Verhaltensweisen in die Berechnungen zu einer Komplexität führt, welche sich nicht oder nur ungenügend genau in einer Faktorenstruktur abbilden lässt. Auf der Suche nach einer möglichen Ursache für dieses Phänomen überprüfe ich abschliessend, ob sich die Wertequadrate auch beim likert-skalierten Format in der Datenmatrix abbilden lassen, wie dies beim Forced-Choice-Format der Fall ist (siehe Anhänge 7.3 bis 7.5). Dazu muss ich auf Grund der Möglichkeit, jede Verhaltensalternative unabhängig einzustufen, ein anderes als das zur Überprüfung bei der Forced-Choice-Variante eingesetzte Vorgehen wählen. Auch die üblicherweise zur Bestimmung der Itemgüte durchgeführte Berechnung der Trennschärfe der Wertequadranten erlaubt noch keine Aussage darüber, ob das Item die Logik des Wertequadrates erfüllt. Versuche mit anderen Berechnungsmöglichkeiten führten schnell zur Erkenntnis, dass in diesem Fall die Überprüfung der Wertequadrate nicht mehr mit einer Berechnung über alle Probanden erfolgen kann. So führe ich diese auf Personenebene durch, was bedeutet, dass ich für jeden Probanden dessen Antwortverteilung in jedem Item und zusammengefasst in jeder der drei Skalen auf die Übereinstimmung mit der Logik des Wertequadrates überprüfe.

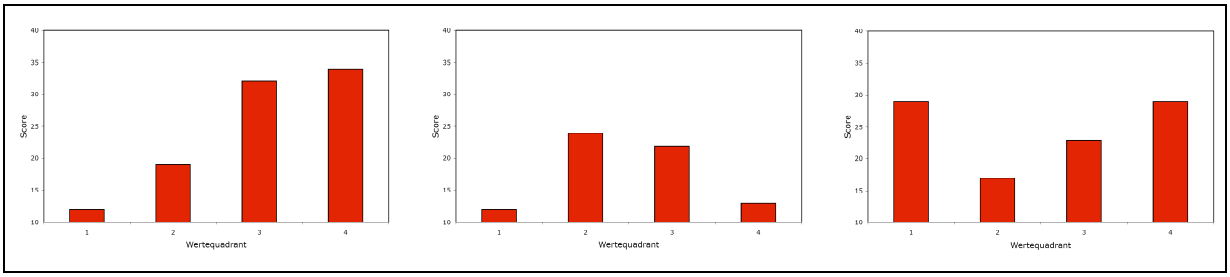


Abbildung 7.8 Beispielhafte Score-Verteilungen in den Wertequadranten auf Skalenebene.

In Abbildung 7.8 habe ich die Antwortverteilungen auf Skalenebene von drei exemplarisch ausgewählten Stellungspflichtigen abgebildet: Die beiden ersten Grafiken zeigen Verteilungen, bei welchen eine Präferenz für einen einzelnen oder zwei benachbarte Wertequadranten feststellbar ist und somit eine eindeutige Interpretation der Antworten auf der betreffenden Skala ermöglicht. Die dritte Grafik zeigt ein widersprüchliches oder inkonsistentes Antwortverhalten, welches zu Präferenzen in nicht-benachbarten Wertequadranten führt. Diese Antwortverteilung lässt sich auch mit der Logik des Wertequadrates nicht sinnvoll interpre-

tieren. Die Aussage, dass das Wertequadrat grundsätzlich eine Erklärung für scheinbar widersprüchliches Verhalten liefert, bezieht sich denn auch nur auf die beiden Tugenden und nicht darauf, dass eine Person gleichzeitig oder zeitnah Verhalten aus beiden Übertreibungen zeigt (z. B. Eberle & Hartwich, 1995; Gloor, 2007). Nach der dem Wertequadrat zugrunde liegenden Theorie ist es ja gerade so, dass eine gestörte Balance zwischen den beiden Tugenden dazu führt, dass das Verhalten in *eine* der Übertreibungen kippt (Helwig, 1948).

Um die Übereinstimmung mit der Logik des Wertequadrates zu überprüfen, untersuche ich, ob die individuellen Antwortverteilungen über die vier Wertequadranten stetig steigend oder fallen sind, respektive einer umgekehrten U-Verteilung folgen. Dazu verwende ich den Datensatz mit den anhand der Trennschärfeanalyse umgepolten Wertequadranten 1 respektive 2. Im Anhang 7.12 sind die Werte aus dieser Überprüfung auf Itemebene aufgeführt. Auf Grund der relativ groben, vierstufigen Antwortskala erfüllt das Antwortverhalten bei 40% der Probanden die Wertequadrat-Logik nicht. Aus diesem Grund führe ich die Analyse auch noch auf Skalenebene durch. Die Ergebnisse dazu habe ich in Tabelle 7.20 dargestellt. Durchschnittlich und über alle drei Skalen hinweg haben 4.0% der Stellungspflichtigen die Höchstausprägung auf dem Wertequadranten 1 respektive den Wertequadranten 1 und 2, 89.8% auf einem oder beiden mittleren Wertequadranten und 2.2% auf dem Wertequadranten 4 respektive den Wertequadranten 3 und 4. Bei 4.0% der Fälle zeigte sich keine eindeutige Präferenz. Dieses Ergebnis lässt den Schluss zu, dass auch beim likert-skalierten Antwortformat die Wertequadrate bei den meisten Personen funktionieren, jedoch nur aggregiert auf Skalenniveau. Auf Itemebene sind einerseits grosse Unterschiede zwischen den einzelnen Items auszumachen – der Anteil uneindeutiger Antwortmuster reicht von 16% bis 63% – andererseits auch zwischen den Stellungspflichtigen mit minimal zwei und maximal 28 uneindeutigen Antwortverteilungen auf 30 Items (M = 11.96, SD = 4.15).

Tabelle 7.20
Überprüfung der Wertequadrate in der likert-skalierten Version des Leadership-Fragebogens

	Anteil der Probanden mit jeweils höchster Antwortausprägung im betreffenden Wertequadranten (WQ)							
	WQ 1	WQ 1-2	WQ 2	WQ 2-3	WQ 3	WQ 3-4	WQ 4	uneindeutig
Durchsetzungsfähigkeit	2.4	1.1	53.2	6.5	29.8	.6	.9	**5.6**
Kontaktfähigkeit	3.9	.7	40.0	4.5	43.9	1.3	1.9	**3.8**
Verantwortungsbewusstsein	3.2	.7	33.1	8.4	50.0	.6	1.3	**2.7**

Anmerkung. N = 1'017.

Zusammenfassend lässt sich festhalten, dass sich mit auf dem Wertequadrat basierenden Scoring-Varianten bei der Forced-Choice-Version des Leadership-Fragebogens keine Verbesserung der Reliabilitäten der drei Leadership-Skalen erzielen lässt. Mit der Einführung einer likert-skalierten Einstufung aller vier Verhaltensweisen pro Item gelingt – wie erwartet – eine deutliche Verbesserung der Reliabilitäten, so dass auch diejenige der Skala Durchsetzungsfähigkeit über .80 liegt, was gemäss der Richtlinie der *European Federation of Psychologists' Associations* (EFPA; Lindley, Bartram & Kennedy, 2008) als gut zu bezeichnen ist. Da sich mit dem likert-skalierten Itemformat die Itemzahl pro Skala von zehn auf 40 erhöht, wäre es – nur das Kriterium der Reliabilität berücksichtigend – sogar möglich, die Skalen von zehn Itemstämmen auf sechs zu reduzieren und damit die Testbearbeitungszeit um 25% zu verkürzen.

Die Likert-Skalierung führt jedoch auch dazu, dass sich die drei Skalen faktorenanalytisch nicht mehr replizieren lassen und sich inhaltlich und methodisch begründbare Faktoren bilden. Insbesondere bildet sich ein Faktor, welcher sich hauptsächlich aus Sub-Items der alternativen Verhaltensweise, also dem zweiten Wertequadranten zusammensetzt. Eine mit den Wertequadranten 1, 3 und 4 durchgeführte Faktorenanalyse führt dann schon zu einem deutlich besseren Ergebnis, wobei der Anteil der erklärten Varianz mit 25% immer noch zu tief ist. Eine auf Item-Stamm-Ebene durchgeführte Faktorenanalyse einer auf insgesamt 18 Items reduzierten Test-Version ergibt dann jedoch die erwünschte Struktur, welche – vergleichbar mit derjenigen der Forced-Choice-Version – keine Fehlladungen und insgesamt nur drei Nebenladungen über .30 aufweist. Mit einer erklärten Varianz von knapp 50% ist diese Lösung sogar deutlich besser als diejenige der Forced-Choice-Version, welche dort 30% beträgt. Auch die Überprüfung der Wertequadrate führt zu einem vergleichbaren Ergebnis: Auf Itemebene kann ich diese nicht bestätigen, auf Skalenebene jedoch schon.

Die referierte Studie mit 100 Psychologie-Studierenden zeigte einerseits auf, dass die Forced-Choice- und likert-skalierte Version vergleichbar sind, dass aber die Testsituation und/oder die Bearbeiter zumindest bei der Skala Verantwortungsbewusstsein die Kennwerte beeinflussen.

Durch die Verwendung der Likert-Skalierung steht eine Version des Leadership-Fragebogens zur Verfügung, welche die Anforderungen an die Reliabilität und faktorielle Validität gut erfüllt. Ein erstes Ziel haben wir damit erreicht. Ein weiteres Ziel war die Entwicklung eines Verfahrens, welches 19jährige Jugendliche gut akzeptieren. Nach der Überprüfung des Bekanntheitsgrades der im Item-Stamm geschilderten Situationen gehe ich der Frage nach, ob wir auch dieses Ziel erreicht haben.

7.3 Bekanntheitsgrad der Items des Leadership-Fragebogens

Wie in Kapitel 6.2 beschrieben, setzten wir bei der Generierung der Item-Stämme den Act Frequency Approach ein, um Situationen aus dem Alltagserleben von Jugendlichen zu finden, welche der Erfahrungsrealität der Stellungspflichtigen möglichst nahe kommen. Die so generierten Situationen haben wir für die Bildung der Item-Stämme angepasst und mit zusätzlichen Situationen ergänzt. Es stellt sich nun die Frage, wie gut der Leadership-Fragebogen das als eine Rahmenbedingung für die Testkonstruktion gesetzte Kriterium der Adressatspezifität erfüllt. Um dies zu überprüfen, erstellte ich eine Version des Fragebogens, bei welcher der Testbearbeiter nach jedem Item befragt wird, ob er die in der Ausgangslage geschilderte Situation schon einmal erlebt hat und ob er sich in die Situation hineinversetzen kann. In Abbildung 7.9 ist ein mit diesen beiden Zusatzfragen ergänztes Item des Leadership-Fragebogens dargestellt.

1. Zelten

Mit ein paar Kollegen verbringen Sie eine Woche Ferien auf einem Zeltplatz am Mittelmeer. Es kommen immer mehr Leute und langsam wird es eng.

- ☐ Sie beschliessen, sich einen anderen Zeltplatz zu suchen, wenn es nicht besser wird.
- ☐ Sie haben gerne Leute um sich herum und geniessen die spontanen Bekanntschaften auf dem Zeltplatz.
- ☐ Sie brauchen auch mal Ihre Ruhe. Deshalb ziehen Sie sich öfters an einen ruhigeren Ort zurück.
- ☐ Sie geniessen den Rummel: Je mehr Leute um Sie herum, desto besser. Sonst wäre es ja langweilig.

Haben Sie eine solche oder ähnliche Situation, wie sie in der Ausgangslage geschildert ist, schon einmal erlebt?	ja ☐	nein ☐
Können Sie sich in diese Situation hineinversetzen?	ja ☐	nein ☐

Abbildung 7.9 Mit den beiden Zusatzfragen zur Bekanntheit der geschilderten Situation ergänztes Item des Leadership-Fragebogens.

Da der Aufwand für die Vorgabe dieser um jeweils zwei Zusatzfragen pro Item erweiterten Version des Leadership-Fragebogens in den Rekrutierungszentren sehr gross gewesen wäre – es hätte ein neuer Test programmiert und ins Testsystem implementiert werden müssen – legte ich ein Papier-und-Bleistift-Testheft, welches die Forced-Choice- und die likert-skalierte Version des Leadership-Fragebogens und den NEO-PI-R umfasste, Studierenden der Psychologie vor,

welche für ihren Einsatz eine Versuchspersonenstunde (im Studium zu erbringende Leistung) gutgeschrieben bekamen. Insgesamt haben 102 Studierende den Persönlichkeitsfragebogen ausgefüllt. Zwei davon schloss ich auf Grund ihres Alters (39 resp. 47 Jahre), welches deutlich von demjenigen der intendierten Testpopulation abweicht, aus der Stichprobe aus. Das Alter der verbleibenden 100 Personen liegt zwischen 18 und 33 Jahren mit einem Durchschnitt von 22.65 Jahren (*SD* = 3.69 Jahre). Der Anteil der Männer in der Stichprobe beträgt 51%.

Da sich sowohl die Alters- wie auch die Geschlechtsverteilung dieser Stichprobe von Studierenden deutlich von denjenigen bei den Stellungspflichtigen unterscheiden – das durchschnittliche Alter der Stellungspflichtigen beträgt knapp 20 Jahre und es nehmen praktisch ausschliesslich Männer an der Rekrutierung teil[2] –, ist abzuklären, ob ein Zusammenhang zwischen den beiden Variablen auf den Grad der Bekanntheit der in den Itemstämmen geschilderten Situationen besteht. Das Alter korreliert mit den Antworten zur Frage danach, ob die Situation schon einmal erlebt wurde mit $r = .37^{**}$, mit derjenigen nach dem sich Hineinversetzen in die Situation mit $r = .16$ (ns). Dass das Alter mit dem Grad der Bekanntheit der geschilderten Situationen in einem Zusammenhang steht, ist leicht nachzuvollziehen, da mit zunehmendem Alter auch die Erfahrungen mit unterschiedlichsten Situationen zunehmen.

Der zur Bestimmung des Einflusses des Geschlechts berechnete t-Test für unabhängige Stichproben ergibt keine signifikanten Unterschiede in den Einstufungen der beiden Zusatzfragen ($t(96) = 1.09$, $p = .28$ resp. $t(96) = -.38$, $p = .71$). Dieses Ergebnis lässt sich darauf zurückführen, dass wir zur Generierung der Item-Stämme mittels des AFA zwei gemischtgeschlechtliche Schulklassen herangezogen haben, welchen wir den Auftrag gaben, sich je eine weibliche und eine männliche Person vorzustellen, was verhinderte, dass die in den Items geschilderten Situationen deutlich näher bei der Erfahrungsrealität eines Geschlechts liegen. Um die Vergleichbarkeit dieser Stichprobe mit den Stellungspflichtigen best möglich zu gewährleisten, habe ich trotzdem nur Männer im Alter zwischen 18 bis 24 Jahren (*M* = 20.11 Jahre, *SD* = 1.76 Jahre) in die endgültige Stichprobe (*n* = 35) aufgenommen. In der nachfolgenden Tabelle 7.21 sind die Mittelwerte der Einstufungen der beiden Zusatzfragen zur Bekanntheit der Items aufgeführt. Der Durchschnittswert bei der Frage nach dem Erleben der Situation beträgt *M* = .48 (*SD* = .25; Range: .03 – .97), derjenige bei der Frage nach dem sich in die Situation Hineinversetzen können *M* = .92 (*SD* = .07; Range: .71 – 1.00). In Anhang 7.13 sind die Werte der Gesamtstichprobe dargestellt.

---

[2] *M* = 19.74 Jahre; Range 17 – 31 Jahre, wobei 99.95% der Stellungspflichtigen zwischen 18 und 25 Jahren alt sind. 99.6% der Stellungspflichtigen sind Männer. Rekrutierungsdatensatz 2005/ 2006, *N* = 59'211.

Tabelle 7.21

Bekanntheitsgrad der Aussagen im Leadership-Fragebogen

Dimension	Item	r_{it}	Situation erlebt		Hineinversetzen	
			M	*SD*	*M*	*SD*
Durchsetzungsfähigkeit	Disco	.20	.66	.48	.94	.24
	Lohnerhöhung	.23	.20	.41	.89	.32
	Fahrer	.25	.63	.49	1.00	.00
	Unterbruch	.26	.97	.17	1.00	.00
	Zugreise	.14	.60	.50	.86	.36
	Aufräumen	.27	.46	.51	.91	.28
	Waschküche	.26	.14	.36	.86	.36
	Geschirr	.23	.23	.43	.89	.32
	Musik	.20	.34	.48	.97	.17
	Probleme	.21	.54	.51	.89	.32
	Auswärts	.21	.57	.50	.89	.32
	Arbeit	.22	.71	.46	.97	.17
	Schülerzeitung	.19	.17	.38	.71	.46
Mittelwert Durchsetzungsfähigkeit			.48	.18	.91	.12
Kontaktfähigkeit	Zelten	.36	.54	.51	.94	.24
	Nachbarn	.53	.29	.46	1.00	.00
	Zugfahrt	.61	.83	.38	.94	.24
	Kurs	.56	.74	.44	1.00	.00
	Schultag	.49	.89	.32	.91	.28
	Flugzeug	.57	.63	.49	.94	.24
	Barmann	.45	.34	.48	.89	.32
	Begleitung	.52	.80	.41	1.00	.00
	Party	.40	.69	.47	.91	.28
	Allein	.44	.69	.47	1.00	.00
	Umzug	.48	.37	.49	.94	.24
	Geburtstag	.52	.83	.38	1.00	.00
	Fitness	.53	.66	.48	1.00	.00
Mittelwert Kontaktfähigkeit			.64	.18	.96	.06
Verantwortungsbewusstsein	Schanze	.30	.34	.48	.94	.24
	Unstimmigkeiten	.35	.43	.50	.91	.28
	Beratungsstelle	.35	.17	.38	.83	.38
	Silvester	.41	.31	.47	.89	.32
	Subvention	.37	.14	.36	.83	.38
	Nachhilfestunden	.41	.23	.43	.91	.28
	Kind	.37	.29	.46	.83	.38
	Malediven	.23	.03	.17	.80	.41
	Wohnung	.35	.37	.49	.89	.32
	Bergtour	.40	.11	.32	.89	.32
	Ampel	.43	.80	.41	1.00	.00
	Autofahren	.39	.60	.50	.94	.24
	Mädchen	.32	.29	.46	.91	.29
Mittelwert Verantwortungsbewusstsein			.32	.17	.89	.15

Anmerkung. n = 35; N r_{it} = 7'871.

Der Bekanntheitsgrad der 39 Item-Stämme fällt weniger hoch aus, als erwartet: Bei der Hälfte liegt er unter 50%, bei einem Fünftel sogar bei 20% oder darunter. Hingegen können sich nur in eine Situation – diejenige mit der Schülerzeitungs-Redaktionssitzung – weniger als 80% der Studienteilnehmer hineinversetzen. Der durchschnittliche Bekanntheitsgrad beträgt über alle 39 Situationen 47.8% und liegt damit leicht über der durchschnittlichen Basisrate von 40.2% der 31 Acts der Skala Risikobereitschaft von Krüger und Amelang (1995). Diese sahen sich auch mit extrem tiefen Basisraten in einzelnen Items konfrontiert. So erreichten folgende, für Risikobereitschaft hochprototypische Situationen, Werte von nur gerade 3% bis 8%: „Im Dschungel übernachtete ich unter freiem Himmel.", „Trotz einer starken Konkurrenz eröffnete ich ein Geschäft.", „Ich sagte mich von der Gruppe los und ging allein durch die Wüste.", „Ich nahm an einer Sitzblockade gegen den Krieg teil.", „Ohne mich sonderlich gut in der Region ausgekannt zu haben, wanderte ich ohne Taschenlampe im Dunkeln durch das Hochgebirge." (Krüger & Amelang, 1995, S. 43 resp. S. 51). Dass tiefe Zustimmungsraten mit zum Teil aussergewöhnlichen Situationen korrespondieren, zeigt sich auch beim Leadership-Fragebogen: Nur ein Proband gab an, eine Freundin gehabt zu haben, welche für die Bezahlung der Ferien auf den Malediven einen Kredit aufnehmen wollte, 11% erlebten schon einmal, dass eine schlecht ausgerüstete Person eine Bergwanderung unternehmen wollte, 14% ärgerten sich darüber, dass ein Nachbar die Waschküche an ihrem Waschtag belegt hatte und 14% machten eine Person darauf aufmerksam, dass sie staatliche Subventionen beanspruchen kann. Hohe Zustimmungsraten erhalten erwartungsgemäss die Schilderungen unspektakulärer, alltäglicher Situationen: 97% wurden schon einmal in einem Gespräch unterbrochen, 83% unternahmen eine längere Reise im Zug, 83% hielten sich an einer Geburtstagsparty mit vielen unbekannten Leuten auf und 80% beobachteten, wie eine Person den Fussgängerstreifen in Anwesenheit eines Kindes bei rot überquerte. Anhand dieser Beispiele lässt sich zudem gut aufzeigen, dass es sich bei der Prototypizität der Situationen und deren Bekanntheitsgrad um zwei weitgehend voneinander unabhängige Merkmale eines Item-Stammes handelt.

Grosse Unterschiede im Bekanntheitsgrad zeigten sich nicht nur auf Item- sondern auch auf Skalenebene: So beträgt dieser im Durchschnitt bei der Skala Kontaktfähigkeit 64%, bei Durchsetzungsfähigkeit 48% und nur gerade 32% bei Verantwortungsbewusstsein. Da die Items der Skala Kontaktfähigkeit in allen bisher durchgeführten Analysen sehr gute Kennwerte erzielten, stellt sich die Frage, ob die Höhe des Bekanntheitsgrades einen Einfluss auf die Güte des Items hat. Dazu habe ich die Korrelationen zwischen den Item-Trennschärfen und den beiden Kennwerten des Bekanntheitsgrades berechnet und in Tabelle 7.22 respekti-

ve Abbildung 7.10 dargestellt. Die Trennschärfen habe ich anhand des Datensatzes aus den Rekrutierungszentren von 2003 berechnet (siehe Tabellen 6.17 bis 6.19). Zusätzlich habe ich noch die durchschnittlichen Bekanntheitswerte („Situation schon einmal erlebt") der sechs Items, welche in der kürzesten Version der likert-skalierten Fragebogenvariante verblieben sind, mit denjenigen der sieben ausgeschiedenen Items verglichen.

Tabelle 7.22
Korrelation zwischen der Trennschärfe der Items und deren Bekanntheitsgrad

Dimension	Korrelation Trennschärfe - Bekanntheitsgrad			durchschnittlicher Bekanntheitsgrad	
	Situation erlebt	hinein-versetzen	erlebt und hinein-versetzen	verblei-bende Items (6)	ausge-schiedene Items (7)
Durchsetzungsfähigkeit	.02	.38	.08	.58	.39
Kontaktfähigkeit	.28	.36	.32	.63	.65
Verantwortungsbewusstsein	.48	.48	.47	.38	.27
Gesamttest	.33*	.44**	.37*	.53	.44

Anmerkung. Korrelationen: n = 13 respektive N = 39; Bekanntheitsgrad: n = 35.

Alle in Tabelle 7.22 aufgeführten Berechnungen deuten auf einen Zusammenhang zwischen dem Bekanntheitsgrad des Item-Stammes und der Trennschärfe des Items hin, wobei es zwischen den drei Skalen grosse Unterschiede in der Höhe der Korrelation gibt. In Abbildung 7.11 habe ich aus diesem Grund die Zusammenhänge noch nach Skalen getrennt dargestellt.

Da ich diese Berechnungen anhand der Daten einer kleinen und spezifischen Stichprobe vorgenommen habe, haben die nachfolgenden Interpretationen hypothetischen Charakter. Auf Grund der Neuartigkeit der Befunde würde es sich lohnen, die hier referierten Ergebnisse nochmals anhand eines umfangreichen Datensatzes zu überprüfen.

Die Ergebnisse dieser Untersuchung führen zum Schluss, dass es für die Qualität der Einstufung des eigenen Verhaltens wichtig ist, dass man die geschilderte Situation genau so oder ähnlich schon einmal erlebt hat oder sich zumindest ohne Probleme gedanklich hineinversetzen kann. Für die Konstruktion eines Persönlichkeits-Fragebogens auf der Basis des Act Frequency Approachs müsste man demnach nicht nur die Prototypizität der von den Probanden geschilderten Situationen als Kriterium für die Itemauswahl berücksichtigen, sondern zusätzlich noch deren Bekanntheitsgrad. Dies hat sicherlich zur Konsequenz, dass sich dadurch die Variationsbreite der Verhaltensweisen einschränkt. Da es hier aber im

Gegensatz zu den Studien von Buss und Craik (z. B. 1986) nicht darum geht, die menschliche Persönlichkeit möglichst detailliert und umfassend zu beschreiben, wird diese Verengung der so gemessenen Konstrukte im Rahmen der Personalselektion höchstens geringfügige negative Auswirkungen haben, wenn dies nicht sogar einen Vorteil darstellt. Die Frage nach dem Zusammenhang der Breite von in der Personalselektion verwendeten Konstrukten auf die Validität der jeweiligen Testverfahren untersuchen Forscher seit Jahrzehnten unter dem Stichwort *bandwidth-fidelity dilemma* (Cronbach & Gleser, 1965). Ich werde auf diesen Aspekt noch ausführlich in Kapitel 8.3 eingehen und mich hier auf die Synthese daraus beschränken: Sowohl mit eng als auch mit breit gefassten Persönlichkeitskonstrukten lassen sich Arbeitsleistung vorhersagen (z. B. Barrick & Mount, 2003; Rothstein & Jelly, 2003; Warr, Bartram & Martin, 2005), es zeichnet sich jedoch ab, dass die Erfassung eng definierter Konstrukte zu besseren Kriteriumsvaliditäten führt (Hough & Oswald, 2008; Rothstein & Goffin, 2006), was zum Teil dem in Kapitel 7.2 beschriebenen Reliabilitäts-Validitäts-Dilemma widerspricht.

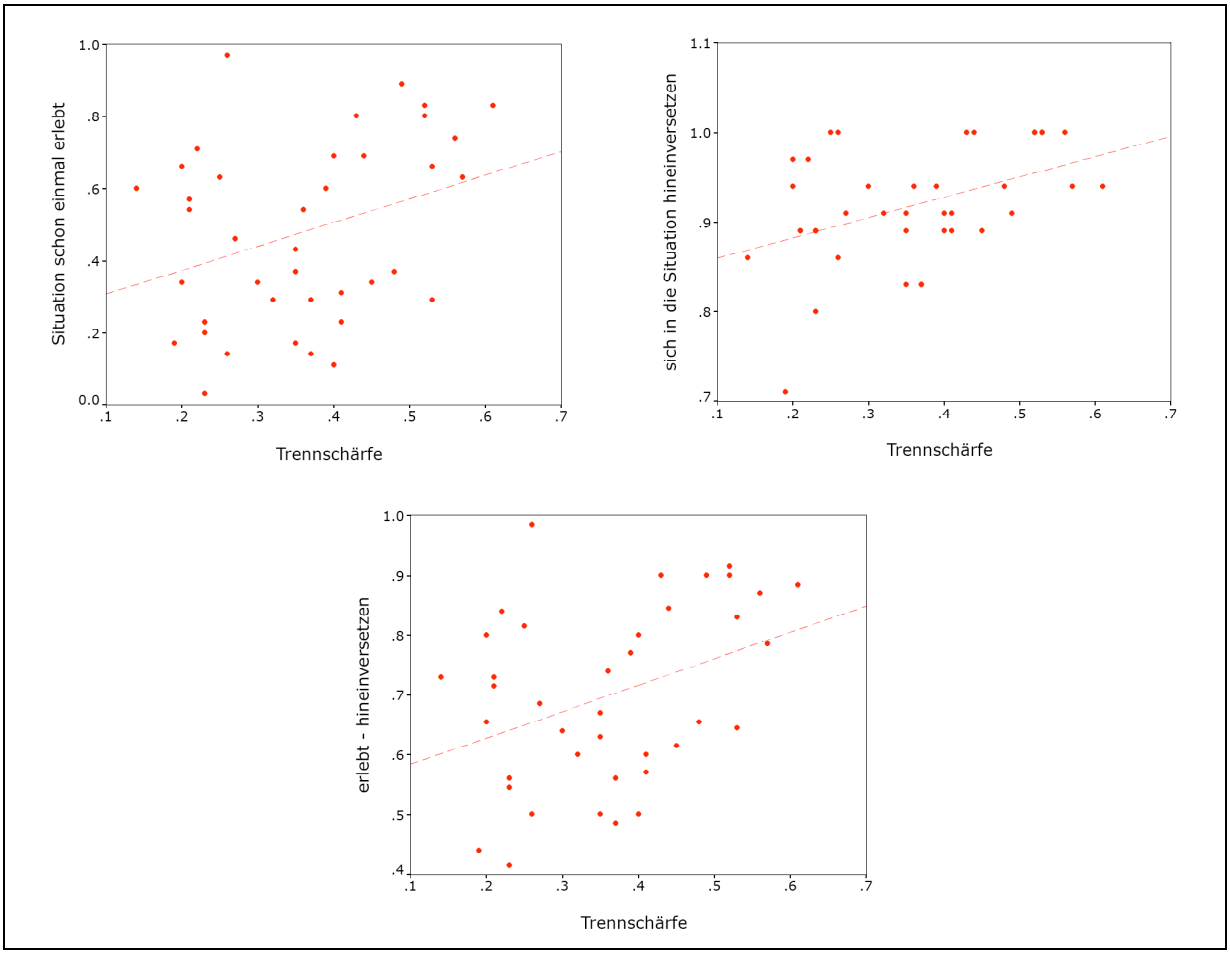


Abbildung 7.10 Grafische Darstellung des Zusammenhangs zwischen dem Bekanntheitsgrad der 39 Item-Stämme und deren Trennschärfen.

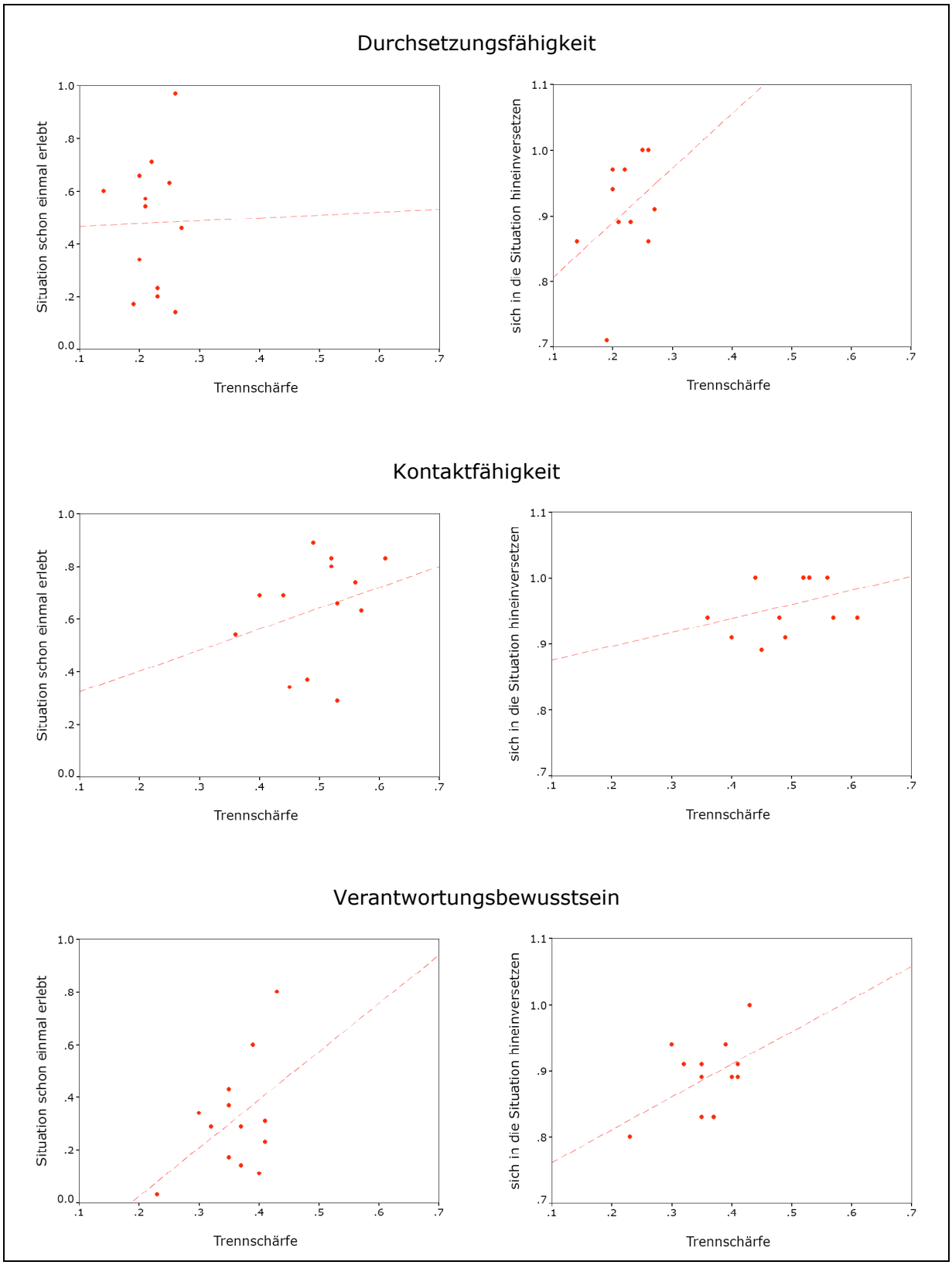


Abbildung 7.11 Grafische Darstellung der Zusammenhänge zwischen dem Bekanntheitsgrad der Items der drei Leadership-Skalen und deren Trennschärfen.

7.4 Studien zur Akzeptanz des Leadership-Fragebogens

Wie schon in der Einleitung erwähnt, war es mir bei der Konstruktion der Testverfahren für die Rekrutierung ein Anliegen, dass diese unter anderem über eine hohe Augenscheinvalidität verfügen, in einer klaren und einfachen Sprache verfasst sind und von den Stellungspflichtigen gut akzeptiert werden (Boss & Baumann, 2003) – Forderungen, welche sich unter den Begriffen Adressatspezifität und Ansprechcharakter subsumieren lassen (Boss, 2005). Um die Akzeptanz des Leadership-Fragebogens zu überprüfen, führte ich drei Studien durch, wobei die dritte in der Realsituation in den Rekrutierungszentren stattfand.

In der ersten Studie verglich eine studentische Arbeitsgruppe (Deiss, Emerson, Imper & Maier, 2002) das Akzeptanzurteil zum Leadership-Fragebogen mit denjenigen von zwei häufig in Persönlichkeitstests eingesetzten Itemformaten: dem likert-skalierten und dem Forced-Choice-Itemformat (siehe Abbildung 7.12). Zur Erfassung der Akzeptanz entwickelten sie ausgehend von den Kriterien für die Konstruktion der Rekrutierungs-Testverfahren, dem Modell der Bewerberreaktionen auf Personalauswahlverfahren von Gilliland (1993) und dem Beitrag von Kersting (1998) zur sozialen Akzeptanz von Testverfahren in der Personalauswahl einen Fragebogen mit insgesamt 15 Items mit einer sechsstufigen Antwortskala mit den Endpunkten „trifft nicht zu" und „trifft voll zu". Dieser umfasst folgende Dimensionen:

- Layout („Das Fragebogenformat empfinde ich ansprechend."),
- Verständlichkeit („Die Aussagen sind für mich verständlich formuliert."),
- Erleben („Das Ausfüllen des Fragebogens hat mir Spass gemacht."),
- Augenscheinvalidität („Ich denke, dass anhand meiner Antworten Aussagen über mein Verhalten gemacht werden können."),
- Alltäglichkeit („Ich konnte mich gut in die beschriebenen Situationen versetzen.") und
- Privatsphäre („Ich fühle mich durch das Beantworten des Fragebogens durchleuchtet.")[3].

Der Test-Fragebogen setzte sich aus je zehn likert-skalierten Items aus einem im Rahmen der Leadership-Testkonstruktion entwickelten Vergleichsfragebogen zu den Dimensionen Verantwortungsbewusstsein und Durchsetzungsfähig-

[3] Alle Items des Fragebogens sind zusammen mit den Item-Kennwerten in Anhang 7.14 aufgeführt.

keit, aus je fünf Leadership-Items zu denselben Dimensionen und 14 Forced-Choice-Tetraden aus einem kommerziell vertriebenen Persönlichkeitsinventar zusammen. Die drei Itemformate sind in Abbildung 7.12 dargestellt. Zur besseren Unterscheidbarkeit haben wir diese auf unterschiedlich eingefärbtes Papier gedruckt. Nach der Bearbeitung des Test-Fragebogens zu den drei Itemformaten hatten die Studienteilnehmer zu jedem Itemformat einen Akzeptanzfragebogen zu beantworten, welche wiederum auf eingefärbtem Papier gedruckt waren. Den Abschluss der Befragung bildete ein Akzeptanz-Vergleichs-Fragebogen, bei welchem die Studienteilnehmer zu sechs Fragen die drei Itemformate rangieren mussten (siehe Tabelle 7.23 weiter unten).

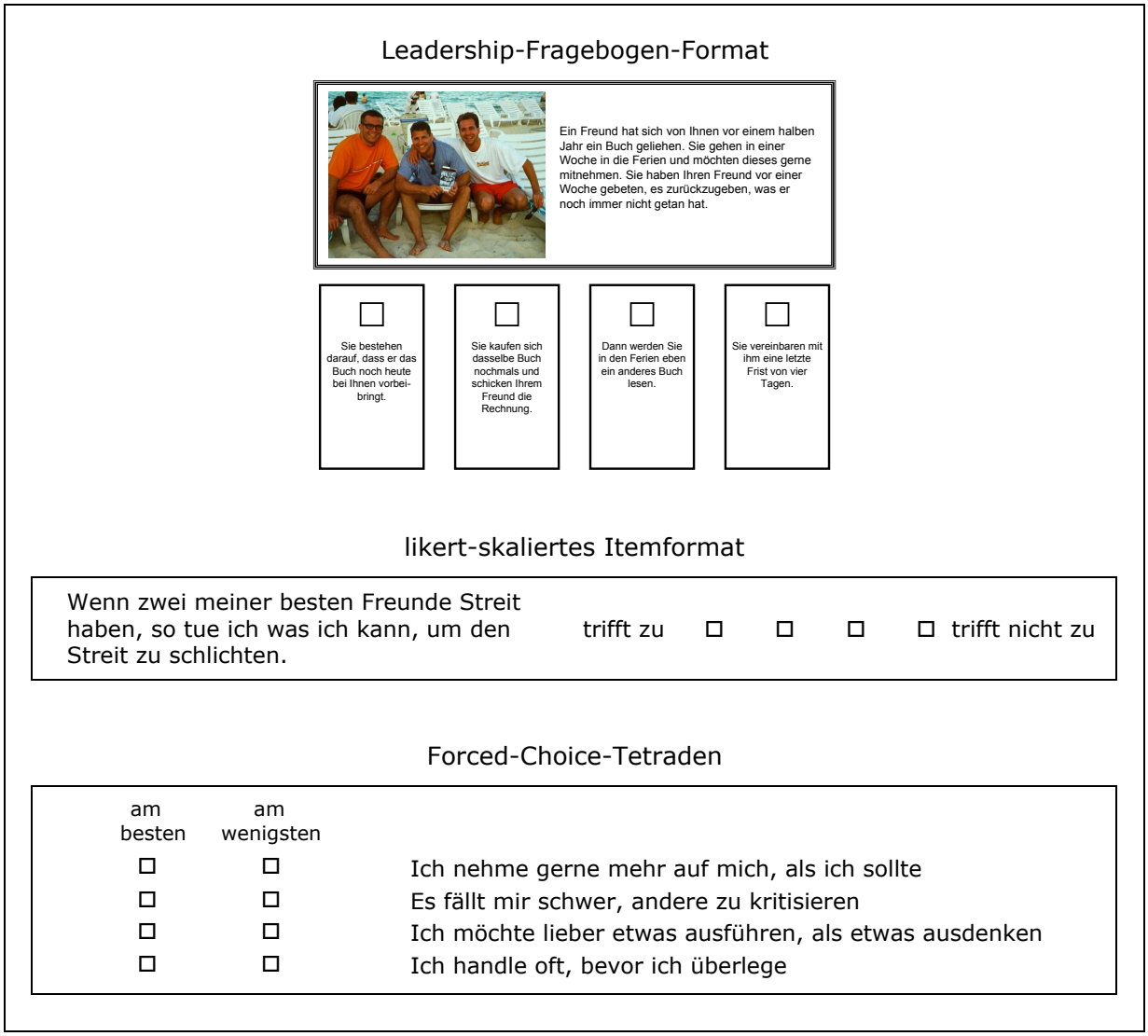


Abbildung 7.12 Itemformate im Test-Fragebogen 1.

Tabelle 7.23
Skalenkennwerte und Interkorrelationen des Akzeptanz-Fragebogens 1

	M	*SD*	1	2	3	4	5
1. Layout	3.98	1.42					
2. Verständlichkeit	4.58	1.29	.45*				
3. Erleben	4.02	1.22	.72*	.47*			
4. Augenscheinvalidität	4.06	.95	.40*	.30*	.48*		
5. Alltäglichkeit	4.15	1.12	.64*	.59*	.65*	.48*	
6. Privatsphäre	4.60	1.17	.12	.35*	.26*	.07	.15

Anmerkung. N = 393. *Bonferroni-korrigiertes Signifikanzniveau $p < .003$.

Die Stichprobe (N = 131) rekrutierte die studentische Arbeitsgruppe in einer Berufsschule (n = 42), einem Gymnasium (n = 39), einer Rekrutenschule (n = 30) und einer Unteroffiziersschule (n = 20). Der Männeranteil beträgt 66.4%. Die Reliabilitäten nach Cronbach Alpha der sechs Skalen sind in Tabelle 7.24 aufgeführt und liegen zwischen α = .60 (Augenscheinvalidität) und α = .82 (Layout). In dieser Tabelle sind zudem die Mittelwerte der Akzeptanzeinstufungen der drei Item-Formate aufgeführt, wobei die höchsten Ausprägungen jeweils fett gedruckt sind[4]. In den Tabellen im Anhang 7.15a sind die Kennwerte einfaktorieller Varianzanalysen mit den entsprechenden Post-hoc-Vergleichen dargestellt. Die Studienteilnehmer haben dem Leadership-Fragebogen-Format in allen Dimensionen mit Ausnahme der Dimension Augenscheinvalidität die höchste Akzeptanz-Einstufung gegeben. Am schlechtesten haben sie die Forced-Choice-Tetraden beurteilt. Der deutlichste Unterschied zwischen den drei Formaten zeigt sich in der Einstufung des Layouts mit den Werten M_{Leader} = 4.79 (SD = 1.25), M_{Likert} = 4.03 (SD = 1.06), $M_{F\text{-}C}$ = 3.11 (SD = 1.41) ($F(2, 256.29) = 51.19$, $p < .001$, $\omega = .48$). Keine signifikanten Unterschiede zwischen den drei Itemformaten ergeben sich in der Dimension Privatsphäre ($F(2, 390) = .95$, $p = .39$, $\omega = .02$). Knapp signifikant wird der Unterschied zwischen dem likert-skalierten Itemformat (M_{Likert} = 4.24, SD = .92) und dem Leadership-Fragebogen-Format (M_{Leader} = 3.97, SD = .90) bei der Beurteilung der Augenscheinvalidität ($F(2, 390) = 3.48$, $p < .05$, ω = .11; Post-hoc-Vergleich (Tukey) $p = .05$).

[4] Sind zwei oder drei der Werte fett gedruckt, so unterscheiden sich diese nicht signifikant voneinander.

Tabelle 7.24

Reliabilitäten der Skalen des Akzeptanz-Fragebogens und durchschnittliche Akzeptanzeinstufungen der drei Itemformate

			Item-Format					
			Leadership		Likert		Forced-Choice	
Skala	Anzahl Items	Cronbach Alpha	*M*	*SD*	*M*	*SD*	*M*	*SD*
Layout	2	.82	**4.79**	1.25	4.03	1.06	3.11	1.41
Verständlichkeit	2	.66	**4.96**	1.10	**4.83**	1.15	3.94	1.35
Erleben	3	.73	**4.49**	1.18	4.07	1.08	3.49	1.18
Augenscheinvalidität	3	.60	3.97	.90	**4.24**	.92	**3.98**	1.01
Alltäglichkeit	3	.68	**4.57**	.98	**4.47**	.96	3.42	1.03
Privatsphäre	2	.68	**4.71**	1.19	**4.52**	1.14	**4.56**	1.17

Anmerkung. $N_{Cronbach\ Alpha}$ = 393; $N_{Mittelwerte}$ = 131. Die Antwortskala reicht von 1 = „trifft nicht zu" bis 6 = „trifft voll zu". Die höchsten Akzeptanzeinstufungen pro Dimension sind fett gedruckt. Sind mehrere Mittelwerte fett gedruckt, so unterscheiden sich diese nicht signifikant voneinander.

Der direkte Vergleich zwischen den drei Itemformaten anhand des Akzeptanz-Vergleichs-Fragebogens – hier mussten die Probanden die drei bearbeiteten Fragebogen bezüglich der generellen Präferenz und bezüglich jeder der fünf Dimensionen in eine Rangreihenfolge bringen (die entsprechenden Fragen sind in Tabelle 7.25 abgebildet) – zeigt ein ähnliches Bild: Wiederum beurteilen die Studienteilnehmer die Forced-Choice-Tetraden am schlechtesten und bevorzugen den Leadership-Fragebogen (M_{Leader} = 1.52, SD = .65, M_{Likert} = 1.79, SD = .66, $M_{F\text{-}C}$ = 2.70, SD = .63; $\chi^2(2) = 93.30$, $p < .001$, $T_{Leader\text{-}Likert}$ = 2'864, $p < .017$, r = -.16; $T_{Leader\text{-}F\text{-}C}$ = 682.5, $p < .001$, r = -.52). Auch zeigen sich wie schon beim Akzeptanz-Fragebogen bei der Dimension Layout die deutlichsten Unterschiede mit grossen Effektstärken: M_{Leader} = 1.16 (SD = .43), M_{Likert} = 2.10 (SD = .58), $M_{F\text{-}C}$ = 2.74 (SD = .48) ($\chi^2(2) = 152.85$, $p < .001$, $T_{Leader\text{-}Likert}$ = 882, $p < .001$, r = -.50; $T_{Leader\text{-}F\text{-}C}$ = 147, $p < .001$, r = -.61). Die Augenscheinvalidität (Frage 6) wurde bei allen drei Formaten ungefähr gleich eingestuft ($\chi^2(2) = 2.80$, $p = .25$). In Tabelle 7.25 sind die mittleren Rangplätze und im Anhang 7.15b die durchgeführten Signifikanzüberprüfungen dargestellt.

Tabelle 7.25

Mittlere Rangierungen der drei Itemformate im Akzeptanz-Vergleichs-Fragebogen

	Item-Format					
	Leadership		Likert		Forced-Choice	
	M	*SD*	*M*	*SD*	*M*	*SD*
1. Wenn Sie nochmals an einer Befragung teilnehmen müssten, welches Frageformat würden Sie bevorzugen?	**1.52**	.65	1.79	.66	2.70	.63
2. Welches Fragebogenformat hat Ihnen vom Layout her am besten gefallen?	**1.16**	.43	2.10	.58	2.74	.48
3. Bei welchem Frageformat konnten Sie sich am schnellsten für eine Antwort entscheiden?	1.84	.66	**1.41**	.59	2.75	.54
4. Welcher Fragebogentyp enthielt die am leichtesten zu verstehenden Aussagen?	**1.62**	.68	**1.61**	.62	2.76	.53
5. Welcher Fragebogentyp enthält Aussagen, die Sie am ehesten ansprechen, da Sie sie mit persönlichen Erfahrungen verbinden können?	**1.67**	.72	**1.80**	.68	2.53	.77
6. Mit welchem Frageformat können Ihrer Meinung nach am ehesten Aussagen über die Persönlichkeit eines Menschen gesammelt werden?	**2.07**	.79	**1.88**	.78	**2.05**	.87

Anmerkung. *N* = 122. Die Werte reichen von 1 – 3 (Rangierung). Die tiefste Rangierung pro Dimension ist fett gedruckt. Sind mehrere Mittelwerte fett gedruckt, so unterscheiden sich diese nicht signifikant voneinander.

Diese erste Akzeptanzstudie zusammenfassend lässt sich sagen, dass die Studienteilnehmer den Leadership-Fragebogen gut bis sehr gut akzeptiert haben: 56.7% von ihnen haben angegeben, dass sie diesen den anderen beiden Formaten gegenüber bevorzugen, 34.4% bevorzugen den likert-skalierten Fragebogenteil und nur 9% die Forced-Choice-Tetraden. Letztere haben auch über alle anderen Akzeptanzurteile im Vergleich zu den beiden anderen Formaten am schlechtesten abgeschnitten. Besonders gut kam das Layout des Leadership-Fragebogens bei den Studienteilnehmern an: 86.1% von ihnen gaben an, dass es ihnen im Vergleich zu den anderen beiden Item-Formaten am besten gefallen hat. Dies zeigt sich auch daran, dass das Bearbeiten dieses Fragebogens am angenehmsten war (Dimension Erleben; z. B. „Ich fand es interessant, den Fragebogen zu beantworten."). Auch das Ziel, einen gut verständlichen Fragebogen zu entwickeln, haben wir erreicht: Mit einem Mittelwert von 4.96 auf einer Skala von 1 bis 6 gaben die Probanden hier die höchste Bewertung überhaupt ab. Weniger erfolgreich waren wir bei der Augenscheinvalidität: So erzielte der Leadership-Fragebogen beim Item „Die Aussagen widerspiegeln Anforderungen, die auch im Berufsleben von einer Führungsperson gefordert werden." eine tiefere Einschätzung als die beiden anderen Verfahren (M_{Leader} = 3.75, SD = 1.17, M_{Likert} = 4.41, SD = 1.09, $M_{F\text{-}C}$ = 4.12, SD = 1.19; $F(2, 390) = 11.01$, $p < .001$, $\omega = .22$).

Da sich die für diese erste Akzeptanzstudie organisierte Stichprobe deutlich von der Population der Stellungspflichtigen unterscheidet und um mehr Erkenntnisse über den Einfluss des Layouts – im Speziellen den Einfluss der die geschilderte Situation illustrierenden Fotografie – auf die Einstufung der Akzeptanz des Leadership-Fragebogens zu erhalten, entschloss ich mich, eine zweite Studie durchzuführen. Dazu kürzte ich den für die Akzeptanzstudie 1 erstellten Leadership-Fragebogen auf acht Items und erstellte zusätzlich eine Version davon im Textformat ohne Fotografien. Der likert-skalierte Fragebogen war derselbe wie in der Akzeptanzstudie 1. In Abbildung 7.13 sind die drei Itemformate dargestellt. Den Akzeptanz-Fragebogen passte ich leicht an, indem ich vier der 15 Fragen durch drei neue ersetzte (siehe Anhang 7.16) und beim Akzeptanz-Vergleichs-Fragebogen mussten die Probanden ein Präferenzurteil abgeben. Für diese Studie erstellte ich vier Fragebogenversionen, welche sich jeweils aus einer der beiden Versionen des Leadership-Fragebogens inklusive Akzeptanz-Fragebogen, dem likert-skalierten Persönlichkeits-Fragebogen inklusive Akzeptanz-Fragebogen und dem Akzeptanz-Vergleichs-Fragebogen zusammensetzten. Um Reihenfolgeffekte zu kontrollieren, setzte ich den Leadership-Fragebogen bei der einen Hälfte der Fragebogen an erster Stelle, bei der anderen an zweiter.

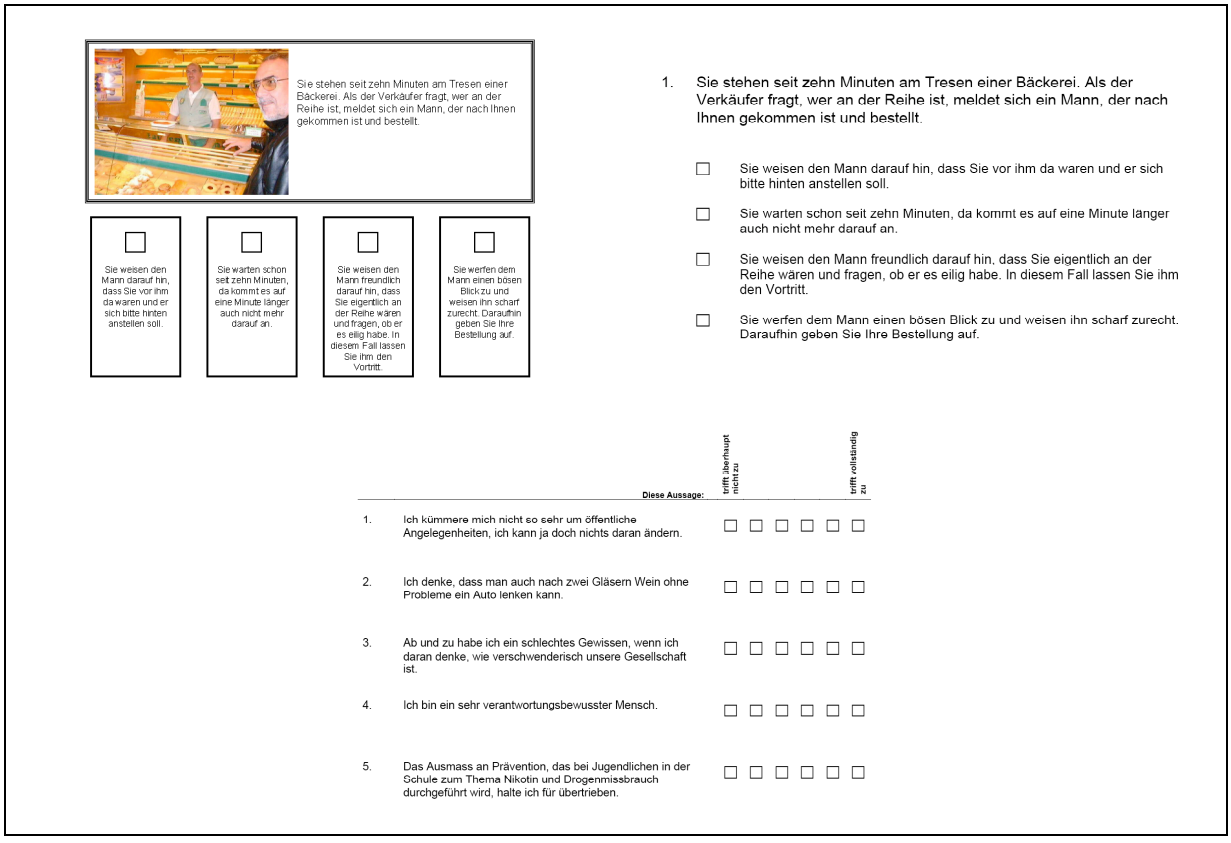


Abbildung 7.13 Itemformate im Test-Fragebogen 2.

Insgesamt nahmen 717 angehende Unteroffiziere der Schweizer Armee an der Studie teil. Einzelne Akzeptanz-Fragebogen schliesse ich von den nachfolgenden Berechnungen aus, da sie unvollständig ausgefüllt sind (siehe Tabelle 7.26). Die Skalenkennwerte und Interkorrelationen habe ich in Tabelle 7.27 dargestellt. Die Unteroffiziers-Schüler stuften im Durchschnitt die Verständlichkeit der Fragebogen-Items auf der sechsstufigen Antwort-Skala am höchsten ein (M = 4.91, SD = .94) und die beiden Fragen nach dem Erleben des Ausfüllens der Fragebogen am tiefsten (M = 3.25, SD = 1.46). Dabei scheint das Interesse beim Ausfüllen der Fragbogen massgeblich durch das Layout beeinflusst zu sein, wie die Korrelation von r = .69 zwischen diesen beiden Skalen belegt.

Tabelle 7.26
Übersicht über die in der Akzeptanzstudie 2 ausgefüllten Fragebogen

	Leadership Bild	Leadership Text	Likert
Anzahl ausgefüllte Akzeptanz-Fragebogen	353	364	717
Anzahl unvollständig ausgefüllte Akzeptanz-Fragebogen	16	21	27
In die Auswertung einbezogene Akzeptanz-Fragebogen	337	343	690

Tabelle 7.27
Skalenkennwerte und Interkorrelationen des Akzeptanz-Fragebogens 2

	M	*SD*	1	2	3	4	5
1. Layout	3.68	1.33					
2. Verständlichkeit	4.91	.94	.22*				
3. Erleben	3.25	1.46	.69*	.14*			
4. Augenscheinvalidität	4.09	1.05	.02	.33*	-.04		
5. Alltäglichkeit	4.16	1.13	.45*	.33*	.40*	.23*	
6. Privatsphäre	4.76	1.08	.16*	.28*	.13*	-.02	.12*

Anmerkung. N = 1'370. *Bonferroni-korrigiertes Signifikanzniveau p < .003.

In Tabelle 7.28 sind die Reliabilitäten (Cronbach Alpha) der Subskalen des Akzeptanz-Fragebogens aufgeführt, welche von α = .56 (Augenscheinvalidität und Privatsphäre) bis α = .88 (Erleben) reichen. Weiter sind pro Skala die Mittelwerte der Akzeptanz der drei Fragebogenversionen aufgeführt, wobei ersichtlich ist, dass die Leadership-Version mit Bild in vier der sechs Akzeptanzdimensionen – die Ausnahmen bilden die Dimensionen Augenscheinvalidität und Privatsphäre – den höchsten Wert erzielt, wobei hier noch anzumerken ist, dass bei der Augenscheinvalidität ein tiefer Wert tendenziell grösserer Fairness entspricht (siehe dazu auch die Ausführungen auf S. 348).

Tabelle 7.28

Reliabilitäten der Akzeptanz-Skalen und Mittelwerte der drei Itemformate

			Item-Format					
			Leadership Bild		Leadership Text		Likert	
Skala	Anzahl Items	Cronbach Alpha	*M*	*SD*	*M*	*SD*	*M*	*SD*
Layout	2	.76	**4.33**	1.14	3.88	1.33	3.27	1.27
Verständlichkeit	3	.58	**5.02**	.85	**4.98**	.92	4.83	.99
Erleben	2	.88	**3.67**	1.42	**3.45**	1.50	2.95	1.40
Augenscheinvalidität	3	.56	**4.07**	1.03	**3.95**	1.04	4.18	1.05
Alltäglichkeit	2	.63	**4.30**	1.08	**4.19**	1.21	4.09	1.11
Privatsphäre	2	.56	**4.82**	1.00	**4.84**	.99	4.68	1.15

Anmerkung. $N_{Cronbach\ Alpha}$ = 1'370; $N_{Leader\ Bild}$ =337; $N_{Leader\ Text}$ = 343; N_{Likert} = 690. Die höchsten Akzeptanzeinstufungen pro Dimension sind fett gedruckt. Sind zwei oder drei der Werte fett gedruckt, so unterscheiden sich diese nicht signifikant voneinander (Bonferroni-korrigiertes Signifikanzniveau $p < .008$). Bei der Augenscheinvalidität entsprechen tiefe Werte einer hohen Akzeptanzeinstufung im Sinne der Fairness.

Auf Grund der Verletzung der Voraussetzungen für die Durchführung einer Varianzanalyse (inhomogene Varianzen und unterschiedlich grosse Stichprobenumfänge) habe ich in einem ersten Schritt mit einem t-Test die beiden Versionen des Leadership-Fragebogens verglichen und in einem zweiten Schritt diese beiden mit dem likert-skalierten Itemformat (siehe Tabelle 7.29). Die Mittelwerte der Einstufungen unterscheiden sich zwischen den beiden Versionen des Leadership-Fragebogens nur beim Layout (M_{Bild} = 4.33, SD = 1.14; M_{Text} = 3.88, SD = 1.33; $t(665.02) = 4.75$, $p < .001$, $r = .18$) signifikant voneinander, wobei die Effektstärke des Unterschiedes klein ist (Cohen, 1992).

Tabelle 7.29

Mittelwertsvergleiche der Akzeptanz zwischen den beiden Versionen des Leadership-Fragebogens und des Leadership-Fragebogens mit dem likert-skalierten Itemformat

	Mittelwertsvergleich zwischen Leadership-Bild und Leadership-Text	Mittelwertsvergleich zwischen Leadership- (Bild & Text) und Likert-Itemformat
Layout	$t(665.02) = 4.75$, $p < .001$, $r = .18$	$t(1'368) = 12.28$, $p < .001$, $r = .32$
Verständlichkeit	$t(678) = .71$, $p = .48$	$t(1'368) = 3.36$, $p < .008$, $r = .09$
Erleben	$t(678) = 1.98$, $p = .05$	$t(1'368) = 7.88$, $p < .001$, $r = .21$
Augenscheinval.	$t(678) = 1.52$, $p = .13$	$t(1'357.27) = -3.10$, $p < .008$, $r = .08$
Alltäglichkeit	$t(672.35) = 1.25$, $p = .21$	$t(1'368) = 2.54$, $p = .01$
Privatsphäre	$t(678) = -.23$, $p = .82$	$t(1'344.70) = 2.69$, $p < .008$, $r = .07$

Anmerkung. N = 680; N = 1'370. Bonferroni-korrigiertes Signifikanzniveau $p < .008$.

Der Vergleich der Akzeptanz-Mittelwerte der beiden Leadership-Fragebogen-Formate zusammengenommen mit dem likert-skalierten Itemformat ergibt ausser für die Dimension Alltäglichkeit signifikante Unterschiede. Am grössten ist die Effektstärke bei der Dimension Layout (t(1'368) = 12.28, p < .001, r = .32). Nur bei der Dimension Augenscheinvalidität schätzen die angehenden Unteroffiziere das likert-skalierte Itemformat höher ein als der Leadership-Fragebogen (t(1'357.27) = -3.10, p < .008, r = .08), wobei der Effekt jedoch klein ist.

Zu leicht anderen Resultaten führt die Auswertung der Daten aus dem Akzeptanz-Vergleichs-Fragebogen[5]: Wie in Tabelle 7.30 und im Anhang 7.17 ersichtlich, präferierten die angehenden Unteroffiziere ausser bei der Frage 3 (Schnelligkeit der Entscheidungsfindung) die beiden Versionen des Leadership-Fragebogens.

Tabelle 7.30
Präferenzurteile im Akzeptanz-Vergleichs-Fragebogen 2

	Item-Format					
	Leadership Bild – Likert (n = 340)		Leadership Text – Likert (n = 347)		Leadership Bild & Text – Likert (N = 687)	
	M	*SD*	*M*	*SD*	*M*	*SD*
1. Wenn Sie nochmals an einer Befragung teilnehmen müssten, welches Frageformat würden Sie bevorzugen?	1.24	.38	1.31	.42	1.28	.40
2. Welches Fragebogenformat hat Ihnen vom Layout her am besten gefallen?	1.15	.33	1.33	.42	1.24	.39
3. Bei welchem Frageformat konnten Sie sich am schnellsten für eine Antwort entscheiden?	1.53	.46	1.52	.46	1.52	.46
4. Welcher Fragebogentyp enthielt die am leichtesten zu verstehenden Aussagen?	1.39	.42	1.41	.42	1.40	.42
5. Welcher Fragebogentyp enthält Aussagen, die Sie am ehesten ansprechen, da Sie sie mit persönlichen Erfahrungen verbinden können?	1.34	.40	1.34	.42	1.34	.41
6. Mit welchem Frageformat können Ihrer Meinung nach am ehesten Aussagen über die Persönlichkeit eines Menschen gesammelt werden?	1.45	.40	1.45	.43	1.45	.42
Durchschnitt	1.35	.25	1.39	.29	1.37	.27

Anmerkung. Die Werte reichen von 1 bis 2: 1 = Präferenz Leadership, 2 = Präferenz Likert, 1.5 = keine Präferenz.

[5] Die Probanden erhielten zum Akzeptanz-Vergleichs-Fragebogen folgende Instruktion: Bei den nachfolgenden Fragen interessiert uns, welcher der beiden Fragebogen Ihnen besser gefallen hat. Bitte entscheiden Sie sich bei jeder Frage für einen Fragebogen und setzen Sie das entsprechende Kreuz. Wenn Sie denken, dass sich die beiden Fragebogen bezüglich einer der sechs Fragen nicht unterscheiden, so wählen Sie die Kategorie „beide gleich". Dies soll aber die Ausnahme darstellen; versuchen Sie sich wenn möglich für einen Fragebogen zu entscheiden.

Die Einstufungen der beiden Versionen des Leadership-Fragebogens – Bild respektive Text – unterscheiden sich nur minimal, ausgenommen bei der Frage 2 nach dem Layout: Hier steigt durch das Hinzufügen der Fotografie die Präferenzrate von 57.1% auf 82.1% (siehe Anhang 7.17). Der Effekt der Fotografie wirkt sich wohl auch auf das Gesamt-Präferenzurteil aus (Frage 1), welches von 62.3% auf 67.7% steigt. Im Anhang 7.17 ist auch ersichtlich, wie deutlich die Reihenfolgeeffekte ausfallen: Zum Beispiel bei der Frage 3 nach der Schnelligkeit der Entscheidungsfindung fällt das Präferenz-Rating bei der Fragebogen-Version Leadership-Text – Likert deutlich zu Gunsten des likert-skalierten Fragebogens aus (Leadership – unentschieden – Likert: 31.6 – 18.1 – 50.3), bei der Fragebogen-Version Likert – Leadership-Text kehrt sich das Verhältnis (Leadership – unentschieden – Likert: 49.4 – 12.5 – 38.1).

Abschliessend habe ich anhand einer schrittweisen Regressionsanalyse untersucht, wie gross der Einfluss der fünf Skalen auf die Skala Erleben („Das Ausfüllen des vorhergehenden Fragebogens hat mir Spass gemacht.“) ist. Die Voraussetzungen für die Durchführung einer Regressionsanalyse sind in diesem Datensatz gegeben: Die Stichprobe ist mit N = 1'370 mehr als ausreichend gross (Field, 2009) und der Datensatz ist frei von Ausreissern, da keines der standardisierten Residuen den Wert 3.29 übersteigt (Tabachnick & Fidell, 2001). Das P-P-Diagramm zeigt auf, dass alle standardisierten Residuen auf der Referenzlinie der Normalverteilung liegen und somit Linearität und Normalverteilung der Residuen gegeben ist (Schendera, 2008). Die Durbin-Watson-Statistik zur Bestimmung der Unabhängigkeit der Residuen beträgt 1.94 und liegt somit innerhalb der Grenzen von 1.5 und 2.5 (Rudolf & Müller, 2004). Wie die Plots *ZRESID gegen *ZPRED zeigen, ist die Homoskedastizität – mit Ausnahme bei der Variablen Privatheit – gegeben. Die Toleranzen der Kollinearitäts-Statistik liegen zwischen .72 und .90, und die Varianzinflationsfaktoren (VIF-Werte) der Prädiktoren im Modell liegen sehr deutlich unter 10, was darauf hindeutet, dass keine Multikollinearität vorliegt (Field, 2009).

Die in Tabelle 7.31 dargestellten Ergebnisse zeigen deutlich auf, dass die Variable Layout („Die Art der Aufmachung der Aufgaben hat mir gut gefallen.“) mit Abstand den höchsten Einfluss auf die Variable Erleben hat, indem sie alleine 47.8% deren Varianz erklärt. Durch das Hinzufügen der Variablen Alltäglichkeit („Ich kenne den Inhalt der Aussagen im Fragebogen aus meinem Alltagsleben“) und Augenscheinvalidität („Dieser Fragebogen ist leicht durchschaubar.“ „Ich kann mir gut vorstellen, welche Persönlichkeitseigenschaften mit diesem Fragebogen untersucht werden.“) ergibt sich nur eine geringfügige zusätzliche Varianzaufklärung von 1.6%. Der negative β-Wert der Variable Augenscheinvalidität zeigt sich auch schon in der Korrelationsmatrix (Tabelle 7.27) und lässt sich

inhaltlich erklären: Ein Fragebogen, welcher leicht durchschaubar ist und bei welchem man sich deshalb auch auf einfache Weise besser darstellen kann als man sich wirklich sieht, ist wohl in den Augen vieler Testbearbeiter weder sehr seriös noch fair, weil er in Abhängigkeit der Verfälschungstendenz zu ungleicher Behandlung führen kann. Dieser Aspekt – die Vergleichbarkeit der Durchführung – stellt eine der zehn Regeln im Modell der Bewerberreaktionen auf Personalselektionsverfahren von Gilliland (1993) dar, welcher sich in verschiedenen Studien als Prädiktor für Fairness bestätigen liess (z. B. Dineen, Noe & Wang, 2004; Madigan & Macan, 2005; Ryan, Greguras & Ployhart, 1996).

Tabelle 7.31
Schrittweise Regressionsanalyse der Akzeptanzvariablen auf das Erleben der Fragebogenbearbeitung

		B	*SE B*	β
Schritt 1	(Konstante)	.44	.08	
	Layout	.76	.02	.69*
Schritt 2	(Konstante)	.04	.11	
	Layout	.71	.02	.64*
	Alltäglichkeit	.15	.03	.11*
Schritt 3	(Konstante)	.43	.15	
	Layout	.70	.02	.63*
	Alltäglichkeit	.18	.03	.14*
	Augenscheinvalidität	-.12	.03	-.08*

Anmerkung. N = 1'370. R^2 = .48 für Schritt 1, ΔR^2 = .01 für Schritt 2, ΔR^2 = .01 für Schritt 3, * p < .001

Anhand dieser Analyse und der oben geschilderten Ergebnisse lässt sich klar aufzeigen, dass dem Layout eines Fragebogens eine grosse Bedeutung hinsichtlich des Erlebens bei dessen Bearbeitung zukommt. Dies ist einer der Aspekte, welchem wir bei der Testkonstruktion grosse Aufmerksamkeit geschenkt haben. Als besonders wirkungsvoll hat sich dabei die Fotografie herausgestellt, welche die im Item-Stamm geschilderte Situation illustriert und welche die einzige Dimension darstellt, in welcher sich die Einstufungen der beiden Leadership-Fragebogen-Formate signifikant voneinander unterscheiden. Als zweiter Aspekt hat die Alltäglichkeit, also das Verbinden der Aussagen des Fragebogens mit persönlichen Erfahrungen, einen Einfluss darauf, wie viel Spass einem die Fragebogenbearbeitung bereitet. Mit der Verwendung des Act Frequency Approachs haben wir genau dieses Ziel verfolgt, indem wir Situationen generieren liessen, welche möglichst nahe bei der Erfahrungsrealität von 19jährigen liegen.

Die oben referierten Studien weisen jedoch zwei Nachteile auf: Erstens habe ich sie nicht in einer realen Selektionssituation durchgeführt und zweitens habe ich den Probanden nur Ausschnitte aus den Fragebogen vorgelegt. Es stellt sich nun also die Frage, wie Stellungspflichtige den Leadership-Fragebogen und ein traditioneller, likert-skalierter Persönlichkeits-Fragebogen beurteilen, nachdem sie diese im Rekrutierungszentrum vollständig bearbeitet haben. Dabei setze ich anstelle des an Gilliland (1993) und Kersting (1998) angelehnten Akzeptanz-Fragebogens die Akzept!-Skalen von Kersting (2006, 2008) ein, welche ich in Kapitel 5.6 beschrieben habe. Nachfolgend führe ich die insgesamt acht Skalen des Akzept!-P (für Persönlichkeits-Fragebogen) und des Akzept!-L (für Leistungstests) nochmals auf:

- *Kontrollierbarkeit* („Bei der Bearbeitung der Fragen/Aussagen wusste ich jederzeit, was ich tun muss.")
- *Messqualität* („Der Leadership-Fragebogen ermöglicht es, die Persönlichkeits-Unterschiede, welche zwischen verschiedenen Menschen bestehen, exakt zu messen.")
- *Augenscheinvalidität* („Dass man mit den Fragen/Aussagen geeignete Personen für eine Kaderposition herausfinden kann, ist zu bezweifeln.")
- *Belastungsfreiheit* (Nur Akzept!-L; „Die Bearbeitung der Testaufgaben ist belastend.")
- *Wahrung der Privatsphäre* (Nur Akzept!-P; „Was ich auf solche Fragen/ Aussagen antworte, geht diejenigen, die die Testergebnisse erhalten, nichts an.")
- *Intention zur unverfälschten Antwort* (Nur Akzept!-P; „Obwohl von diesem Leadership-Fragebogen eine wichtige Entscheidung für mich abhängt, habe ich mich nicht anders dargestellt als ich bin.")
- *Antwortfreiheit* (Nur Akzept!-P; „Aufgrund der vorgegebenen Antwortmöglichkeiten hatte ich nicht die Freiheit, so zu antworten, wie es für meine Person zutreffend ist.")
- *Gesamtbeurteilung* („Welche Schulnote würden Sie dem soeben bearbeiteten Leadership-Fragebogen geben?")

In Anhang 7.18 habe ich die beiden eingesetzten Akzept!-Skalen – den Akzept!-P mit 23 Aussagen plus fünf Zusatzfragen und den Akzept!-L mit 16 Aussagen plus acht Zusatzfragen – vollständig aufgeführt. Für den Einsatz der Akzept!-Skalen in den Rekrutierungszentren passten wir die allgemein gehaltenen Aussagen der Originalskalen an die Testsituation und jeweiligen Testver-

fahren an: So ersetzten wir zum Beispiel die Aussage „Um beim Verfahren gut abzuschneiden..." mit der Aussage „Um im Leadership-Fragebogen gut abzuschneiden..." oder ersetzten den Begriff Job durch Militärdienst. Zudem überprüften wir die Verständlichkeit der Aussagen. Dazu führten wir mit Stellungspflichtigen des Rekrutierungszentrums Mels, welche im standardmässig in den Rekrutierungszentren durchgeführten Textverständnistest eine sehr tiefe Punktzahl erreicht hatten (unter Prozentrang 20), einen kognitiven Pretest (Prüfer & Rexroth, 2000) durch. Hierzu legten wir den Stellungspflichtigen jeweils einen Ausschnitt des zu beurteilenden Fragebogens vor und liessen sie anschliessend vier bis fünf Akzept!-Aussagen laut vorlesen. Sie nahmen daraufhin die Einstufung auf der Antwortskala vor, worauf wir die Kontrollfragen zur Überprüfung des Textverständnisses stellten. Dabei setzten wir folgende Techniken ein:

- *Category Selection Probing:* Der Stellungspflichtige begründet, warum er sich für bestimmte Antwortvorgaben/Skalenwerte entschieden hat.
- *Informational Retrieval Probing:* Der Stellungspflichtige beschreibt, welche Gedanken er sich bei der Beantwortung der Frage gemacht hat.
- *General Probing:* Dem Stellungspflichtigen wird eine sehr allgemein gehaltene Zusatzfrage nach dem Verständnis gestellt.
- *Overall Rating:* Im Anschluss an die Bearbeitung des Fragebogens wird dem Stellungspflichtigen die Gelegenheit gegeben, sich rückblickend gesamthaft zum Fragebogen zu äussern.

Wir nahmen die Antworten der Stellungspflichtigen auf Tonband auf und notierten diese zusätzlich stichwortartig. Insgesamt konnten wir sieben Stellungspflichtige mit dem Akzept!-L und zehn mit dem Akzept!-P befragen. Die Auswertung der Notizen und Tonbandaufzeichnungen ergab, dass die befragten Stellungspflichtigen die Mehrheit der Aussagen nicht vollständig richtig verstanden hatten, da sie den Inhalt nicht korrekt wiedergeben konnten. Die detaillierte Auswertung ergab folgende Gründe dafür: Zu lange Sätze, verschachtelte Sätze, Sätze mit Verneinungen und Fremdwörter. Diese Faktoren sind aus der Lesbarkeitsforschung bekannt und sind in die Lesbarkeitsformeln zur Bestimmung der Komplexität von Texten eingeflossen (z. B. Klare, 1963; Tauber, Stoll & Drewek, 1980; siehe auch Boss, 1999). Um unsere Ergebnisse aus dem Einsatz der Akzept!-Skalen mit deren aus anderen Studien vergleichen zu können, habe ich darauf verzichtet, die einzelnen Aussagen sprachlich zu überarbeiten und zu vereinfachen. Im Hinblick auf eine umfassende Überprüfung der Akzeptanz der in den Rekrutierungszentren eingesetzten Testverfahren drängt sich eine Überarbeitung – auch unter Berücksichtigung der nachfolgend präsentierten Ergebnisse – jedoch auf.

Insgesamt haben im Rahmen dieser Akzeptanz-Studie in den Rekrutierungszentren Mels und Rüti 648 Stellungspflichtige jeweils eines der folgenden standardmässig auf Computern durchgeführten Testverfahren beurteilt:

- Den Leadership-Fragebogen in der Version mit 30 Items und Forced-Choice-Antwortformat.
- Den Persönlichkeits-Fragebogen, welcher mit 80 sechsstufig likert-skalierten Items die Dimensionen Leistungsmotivation, Belastbarkeit, Extraversion, Gewissenhaftigkeit, Entgegenkommen/Friedfertigkeit, Teamfähigkeit und sozial erwünschtes Antwortverhalten erfasst.
- Den Intelligenztest 95, bestehend aus einem Verbal- und einem Figuralteil mit je 30 Items.

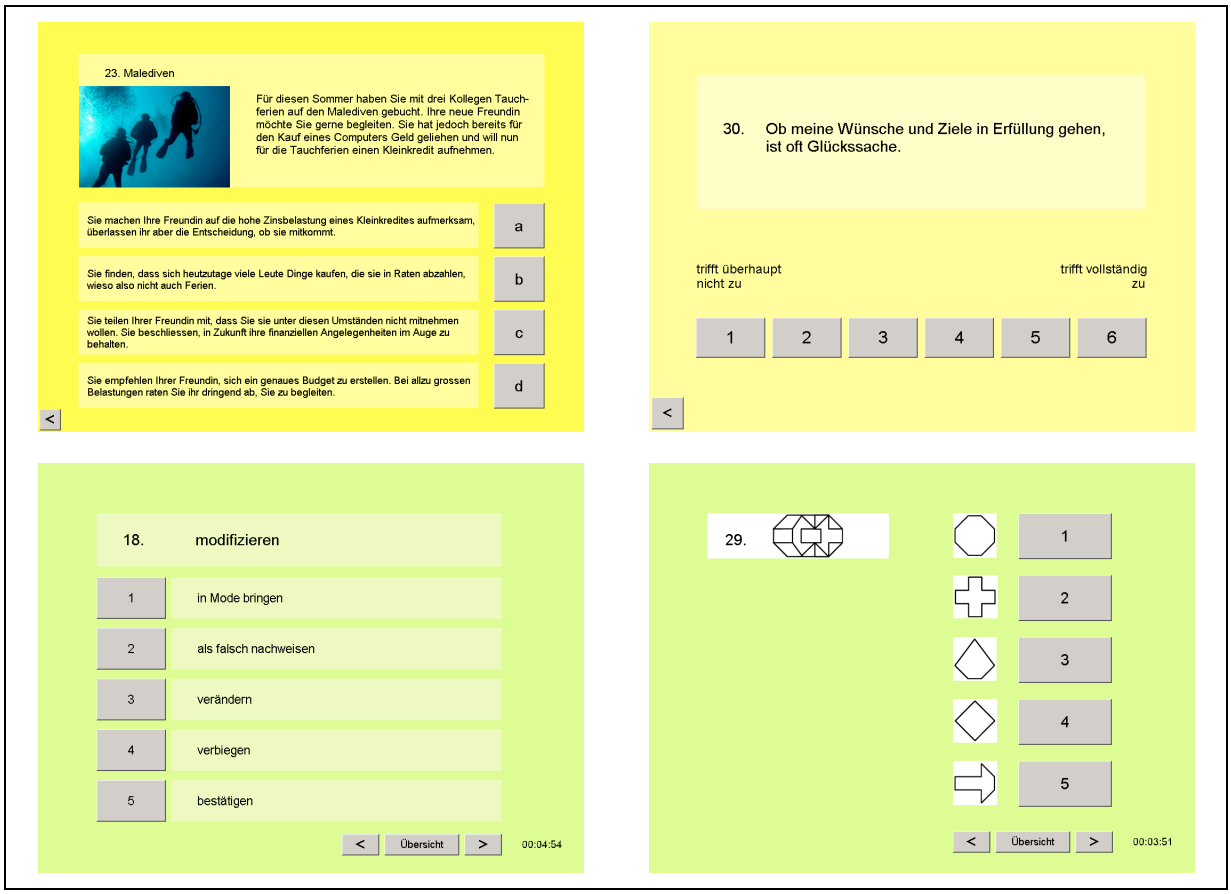


Abbildung 7.14 Layout der drei beurteilten Testverfahren.

Die Stellungspflichtigen wurden dazu angehalten, die im Format A5 gedruckten Akzept!-Fragebogen unmittelbar im Anschluss an die computergestützte Bearbeitung des zu beurteilenden Testverfahrens auszufüllen. Da es sich bei der ersten

Datenerhebung herausstellte, dass viele von ihnen diesen jedoch erst am Ende des jeweiligen Testblockes ausfüllte, schalteten die Testassistentinnen nach dem zu beurteilenden Test ein Pausenbild auf, welches den Stellungspflichtigen signalisierte, dass sie nun den Fragebogen auszufüllen haben. Nach Abzug der unvollständig ausgefüllten Fragebogen verbleiben 606 Fragebogen im Datensatz. In Tabelle 7.32 sind die Teildatensätze aufgeführt.

Tabelle 7.32
Übersicht über die Stichprobe zur Überprüfung der Akzeptanz des Leadership-Fragebogens, des Persönlichkeits-Fragebogens und des Intelligenztests

	Teilstichproben Rekrutierungszentren Mels / Rüti			Gesamtstichprobe		
	ausgefüllt	ungültig	total	ausgefüllt	ungültig	total
Leadership-Fragebogen	113 / 125	11 / 7	102 / 118	238	18	220
Persönlichkeits-Fragebogen	86 / 108	13 / 3	73 / 105	194	16	178
Intelligenztest	104 / 112	8 / 0	96 / 112	216	8	208

Anmerkung. Bei den Teilstichproben des Persönlichkeits-Fragebogens und des Intelligenztests in Rüti fehlen die Angaben zu den Zusatzfragen.

In Anhang 7.20 habe ich die Itemkennwerte und in Anhang 7.21 die Skalenkennwerte, Skalenreliabilitäten und die Skalen-Interkorrelationen der Akzept!-Skalen für jedes der drei beurteilten Testverfahren aufgeführt. Die Reliabilitäten schwanken zwischen α = .49 und .85 und liegen so zum Teil knapp an der Grenze von α = .50 für gruppendifferenziell-diagnostische Aussagen (Rückert, 1993). Bei den Interkorrelationen fällt bei allen drei Testverfahren der hohe Wert zwischen der Skala Messqualität und Augenscheinvalidität auf, welcher zwischen r =.61 und .66 liegt. Die explorativen Faktorenanalysen mit Varimax-Rotation, deren Ergebnisse in Anhang 7.24 dargestellt sind, bestätigen die inhaltliche Verwandtschaft dieser beiden Skalen: Bei den drei Testverfahren laden die acht Items zur Messqualität und zur Augenscheinvalidität auf einen Faktor. Somit wird auch ersichtlich, dass sich die sechs respektive vier Skalen der beiden Akzept!-Versionen faktorenanalytisch nicht bestätigen liessen. Interessant ist, dass sich die Faktorenstrukturen beim Leadership- und beim Persönlichkeits-Fragebogen unterscheiden, obwohl es sich um dieselbe Akzept!-Skala handelt: Der Scree-Plot deutet bei ersterem auf eine Fünf- oder Zwei-Faktoren-Lösung hin, bei letzterem eindeutig auf eine Drei-Faktoren-Lösung. Auf Grund der für die Durchführung einer Faktorenanalyse eher geringen Stichprobenumfängen (n = 220 resp. n = 172) – Comrey und Lee (1992) empfehlen 300 (gut) bis 500 (sehr gut) Probanden und Guadagnoli und Velicer (1988) raten bei weniger als 300 Probanden

davon ab, Faktoren zu interpretieren, welche sich aus schwachen Ladungen zusammensetzen – ist eine Interpretation dieses Befundes jedoch heikel.

Die Skalenwerte der Zusatzfragen sind in Anhang 7.22 dargestellt. Da die Angaben zu den Aussagen „Die Bearbeitung hat mir Spass gemacht." und „Die Bearbeitung war unterhaltsam." hoch korrelieren (zwischen r = .76 und .84) habe ich sie zu einem Wert zusammengefasst. Bei den Zusatzfragen fallen die hohen Interkorrelationen zwischen den beiden Variablen zum Spass an der Bearbeitung des Fragebogens und der Variablen zum Gefallen an der Darstellung auf, welche zwischen r = .56 und .74 liegt. Diesen Befund werde ich weiter unten noch anhand einer Regressionsanalyse erhärten.

Tabelle 7.33
Skalenkennwerte der drei Testverfahren in den Akzeptanzskalen

	Testverfahren					
	Leadership		Persönlichkeit		Intelligenz	
Skala	*M*	*SD*	*M*	*SD*	*M*	*SD*
Kontrollierbarkeit	**5.36**	.76	**5.31**	.87	5.01	.97
Messqualität	**3.65**	.97	**3.81**	1.10	3.30	.98
Augenscheinvalidität	**3.85**	.94	**3.97**	.94	3.13	1.02
Wahrung der Privatsphäre	**4.98**	1.09	**4.81**	.96		
Intention zur unverfälschten Antwort	**4.97**	1.09	**5.02**	.90		
Antwortfreiheit	**4.15**	1.05	**4.25**	1.07		
Gesamtbeurteilung	**4.28**	.88	**4.46**	.73	4.15	.91
Bearbeitung hat Spass gemacht	2.90	1.41	2.94	1.40	**3.30** 2.86	1.49 1.24
Darstellung gefällt mir	**3.41**	1.44	**3.27**	1.49	**3.58** 2.98	1.53 1.44
Konnte mich gut in die Situationen versetzen	**4.35**	1.33	3.92	1.43		

Anmerkung. $N_{Leadership}$ = 220; $N_{Persönlichkeit}$ = 178; $N_{Intelligenz}$ = 208 (resp. N =96 bei den Zusatzfragen). Die höchsten Akzeptanzeinstufungen pro Dimension sind fett gedruckt. Sind mehrere der Werte fett gedruckt, so unterscheiden sich diese nicht signifikant voneinander. Liegen beim Intelligenztest zwei Wertepaare vor, so bezieht sich das erste auf den Verbalteil und das zweite auf den Figuralteil des Tests.

In Tabelle 7.33 ist ersichtlich, dass sich die Akzeptanz-Einstufungen des Leadership-Fragebogens und des Persönlichkeits-Fragebogens sehr ähnlich sind und sich nur bei der Frage nach dem sich Hineinversetzen in die geschilderte Situation signifikant voneinander unterscheiden, wobei der Effekt klein ist (M_{Leader} = 4.35, SD = 1.33, $M_{Persönlichkeit}$ = 3.92, SD = 1.43; $t(290) = 2.38$, $p < .05$, $r = .14$). Die Stellungspflichtigen stufen die beiden Fragebogen jedoch in mehreren Akzeptanz-Dimensionen signifikant höher ein als den Intelligenztest:

Kontrollierbarkeit: M_{Leader} = 5.36, SD = .76, $M_{Persönlichkeit}$ = 5.31, SD = .87, $M_{Intelligenz}$ = 5.01, SD = .97; $F(2, 387.15) = 9.28$, $p < .001$, $\omega = .17$

Messqualität: M_{Leader} = 3.65, SD = .97, $M_{Persönlichkeit}$ = 3.81, SD = 1.10, $M_{Intelligenz}$ = 3.30, SD = .98; $F(2, 601) = 12.71$, $p < .001$, $\omega = .19$

Augenscheinvalidität: M_{Leader} = 3.85, SD = .94, $M_{Persönlichkeit}$ = 3.97, SD = .94, $M_{Intelligenz}$ = 3.13, SD = 1.02; $F(2, 603) = 44.48$, $p < .001$, $\omega = .35$

Zudem stufen die Stellungspflichtigen den Persönlichkeits-Fragebogen in der Gesamtbeurteilung signifikant höher ein, als den Intelligenztest: $M_{Persönlichkeit}$ = 4.46, SD = .73, $M_{Intelligenz}$ = 4.15, SD = .91; $F(2, 573) = 5.94$, $p < .01$, ω = .13; Post-hoc-Vergleich (Tukey) $p < .01$. Diese Ergebnisse zusammenfassend lässt sich sagen, dass die Stellungspflichtigen den Intelligenztest als ein ungenauer messenderes Verfahren als die beiden Persönlichkeitstests beurteilen, obwohl es ein gesicherter Befund ist, dass Intelligenztests zu den validesten Verfahren in der Personalselektion zählen (z. B. Schmidt & Hunter, 1998). Damit zeigt sich hier ein Phänomen, welches mit dem von Cropanzano (1994) beschriebenen *„justice dilemma*" vergleichbar ist: Bewerber schätzen valide Testverfahren eher weniger und stufen sie tendenziell als unfairer ein, als weniger valide Verfahren. Nach Van den Bos, Bruins, Wilke und Dronkert (1999) tritt dies vor allem dann auf, wenn der Bewerber der Überzeugung ist, dass das Testverfahren die Leistungsfähigkeit oder Persönlichkeit genau abbildet und dieses Testergebnis in seinen Augen einen negativen Entscheid im Selektionsprozess mitbeeinflusst. Dies führt dazu, dass der Bewerber das Testverfahren als selbstwertbedrohlich wahrnimmt, was somit in der paradoxen Situation resultiert, dass das Selbstwertgefühl negativ mit der Gerechtigkeitswahrnehmung korreliert (Cropanzano & Wright, 2003). Dies kann jedoch auch damit zusammenhängen, dass die Bewerber das Ergebnis in einem Leistungstest als weniger beeinflussbar erleben, wie sich dies auch in unserer Studie zeigt, indem die Stellungspflichtigen angegeben haben, dass sie beim Intelligenztest weniger genau gewusst haben, was von ihnen verlangt wird und sie die Aufgaben als weniger klar und verständlich aufgefasst haben, als dies beim Leadership- und Persönlichkeits-Fragebogen der Fall war. Zudem schätzen Bewerber Intelligenztests im Allgemeinen weniger als Persönlichkeits-Fragebogen, weil ein negatives Feedback zur wichtigen und umfassenden Eigenschaft Intelligenz bedrohlicher ist, als wenn dies eine eng definierte Persönlichkeitseigenschaft betrifft, welches sie dann auch gut mit selbstwertschützenden Gedanken abschwächen können (Cropanzano & Wright, 2003).

Tabelle 7.34

Schrittweise Regressionsanalyse der Variablen des Akzept!-Fragebogens auf die Gesamtbeurteilung des Leadership-Fragebogens

		B	*SE B*	β
Schritt 1	(Konstante)	3.23	.14	
	Darstellung	.30	.04	.50***
Schritt 2	(Konstante)	2.21	.22	
	Darstellung	.23	.04	.37***
	Augenscheinvalidität	.34	.06	.35***
Schritt 3	(Konstante)	1.91	.22	
	Darstellung	.17	.04	.28***
	Augenscheinvalidität	.28	.06	.29***
	Hineinversetzen	.16	.04	.25***
Schritt 4	(Konstante)	1.73	.23	
	Darstellung	.15	.04	.25*
	Augenscheinvalidität	.17	.07	.18***
	Hineinversetzen	.16	.04	.25***
	Messqualität	.19	.06	.21**
Schritt 5	(Konstante)	1.42	.26	
	Darstellung	.14	.04	.24***
	Augenscheinvalidität	.13	.07	.13
	Hineinversetzen	.14	.04	.22***
	Messqualität	.19	.06	.21**
	Privatsphäre	.11	.05	.14*

Anmerkung. $N = 207$. $R^2 = .25$ für Schritt 1, $\Delta R^2 = .11$ für Schritt 2, $\Delta R^2 = .05$ für Schritt 3, $\Delta R^2 = .03$ für Schritt 4, $\Delta R^2 = .02$ für Schritt 5, * $p < .05$, ** $p < .01$, *** $p < .001$.

Abschliessend untersuche ich, welche Skalen des Akzept!-Fragebogens das Ergebnis in der Gesamtbeurteilung („Welche Schulnote würden Sie dem soeben bearbeiteten Fragebogen geben?") des Leadership-Fragebogens beeinflussen. Tabelle 7.34 stellt die Ergebnisse der durchgeführten schrittweisen Regressionsanalyse dar. Die Voraussetzungen für die Durchführung einer Regressionsanalyse sind auch hier gegeben: Die Stichprobe ist bei zehn Prädiktoren mit $N = 207$ genug gross (Minimum nach Green (1991): 50 + 8x10 = 130) und der Datensatz ist frei von Ausreissern, da keine der standardisierten Residuen den Wert 3.29 übersteigt (Tabachnick & Fidell, 2001). Linearität und Normalverteilung der Residuen sind gegeben, da gemäss P-P-Diagramm alle standardisierten Residuen auf oder in unmittelbarer Nähe der Referenzlinie der Normalverteilung liegen (Schendera, 2008). Die Durbin-Watson-Statistik zur Bestimmung der Unab-

hängigkeit der Residuen beträgt 1.98 und liegt somit innerhalb der Grenzen von 1.5 und 2.5 (Rudolf & Müller, 2004). Die Homoskedastizität ist bei allen Variablen gegeben (Plots *ZRESID gegen *ZPRED). Die Toleranzen der Kollinearitäts-Statistik liegen zwischen .37 und .79 und übersteigen somit den Grenzwert von .20. Zudem liegen auch die Varianzinflationsfaktoren (VIF-Werte) der Prädiktoren im Modell deutlich unter 10, was beides darauf hindeutet, dass keine Multikollinearität vorliegt (Field, 2009).

Als erste Variable wird die Darstellung ins Modell aufgenommen. Sie alleine erklärt jedoch gerade einmal 25.1% der Varianz. Weiter folgen die Variablen Augenscheinvalidität, sich Hineinversetzen, Messqualität und Privatsphäre. Alle fünf Variablen zusammen erklären 44.7% der Varianz der abhängigen Variablen Gesamturteil. Die Ausprägung des Gefallens der Art der Darstellung der einzelnen Situationen wirkt sich demnach stark auf die Gesamtbeurteilung und den Spass an der Bearbeitung des Leadership-Fragebogens aus. Damit ist auch das Ergebnis der Regressionsanalyse mit dem Akzeptanz-Fragebogen 2 bestätigt, bei welchem das Layout den grössten Einflussfaktor auf die Variable Erleben der Testdurchführung hat.

Als wichtigsten Befund der in diesem Kapitel referierten Analysen taxiere ich die mehrfach nachgewiesene Bedeutung des Layouts für das Erleben der Testdurchführung und die Gesamtbeurteilung eines Testverfahrens. Eindrücklich ist dabei der Effekt, welcher sich durch das Hinzufügen einer Fotografie bewirken lässt. Diese Ergebnisse werden jedoch durch die Studie mit den Akzept!-Skalen relativiert, in welcher sich kaum Unterschiede in der Akzeptanzeinstufung des Leadership- und des Persönlichkeits-Fragebogens ergeben haben. Offenbar nehmen die Probanden die Akzeptanz-Einstufungen akzentuierter vor, wenn sie zwei oder mehrere Verfahren gleichzeitig vergleichen müssen. Dies zeigt sich bei der Beurteilung der Darstellung respektive des Layouts: Hier ergeben sich in den Post-hoc-Vergleichen keine signifikanten Unterschiede zwischen den drei Testverfahren, obwohl man hätte annehmen müssen, dass die Stellungspflichtigen den Leadership-Fragebogen deutlich besser beurteilen, als den eher langweilig gestalteten Persönlichkeits-Fragebogen. Interessanterweise zeigt sich jedoch ein Unterschied zwischen den Einstufungen des Verbal- und Figuralteils des Intelligenztests, welche die Stellungspflichtigen gleichzeitig vorgenommen haben. Aber auch dieser Unterschied ist nur schwer nachvollziehbar, da die Darstellung des Figuralteils des Intelligenztests ungewöhnlicher und interessanter ist, als die des Verbalteils, dessen Aufmachung jedoch besser gefiel. Diese Ungereimtheiten deuten darauf hin, dass die Beurteilung der einzelnen Akzeptanz-Aspekte durch die Gesamtbeurteilung des Testverfahrens mitbeeinflusst wird. Da viele Stellungspflichtige den Figuralteil schwieriger als den Verbalteil erleben, könnte es

durchaus sein, dass auch dieser Aspekt in die Beurteilung miteinfliesst. Zudem könnte es sein, dass die Dienst- und Führungsmotivation der Stellungspflichtigen die Testmotivation so stark beeinflusst, dass sich dies auch in der Akzeptanz-Bewertung der Tests niederschlägt. So beurteilen die Stellungspflichtigen die Verfahren nicht mehr nach objektiven Kriterien sondern hauptsächlich durch eine von der subjektiven Motivlage gefärbte Brille. Dieses Phänomen könnte dann auch eine Erklärung für die Unterschiede in den Akzeptanzeinschätzungen des Leadership- und der Persönlichkeits-Fragebogens durch die Stellungspflichtigen und die Unteroffiziers-Schüler liefern.

Aus diesen Befunden und Überlegungen leite ich die Hypothese ab, dass im Realsetting einer Selektionssituation neben dem Ausmass der wahrgenommenen Selbstwertbedrohung auch die Grundeinstellung zur Testung einen massgeblichen Einfluss auf die Akzeptanz-Einstufung der Testverfahren ausübt. Dies ist zum Teil auch im integrativen Modell der Bewerberreaktionen auf Personalauswahlverfahren von Hausknecht, Day und Thomas (2004) so enthalten (siehe Kapitel 5.5): Die wahrgenommenen Prozessmerkmale des Selektionsprozesses, zu welcher auch die wahrgenommene Testschwierigkeit zählt, wirken auf die Wahrnehmungen des Bewerbers, welche Aspekte wie die Gerechtigkeit, Testangst und –motivation und die Einstellung gegenüber den Tests und der Selektion umfasst. In der Diskussion (Kapitel 8) werde ich ein auf der Grundlage der hier referierten Befunde erstelltes Modell zur Erklärung der Akzeptanz von Testverfahren in der Personalselektion vorstellen. Damit gehe ich ganz im Sinne von Anderson (2003, S. 128) vor, welcher der Ansicht ist, dass einfache Akzeptanzuntersuchungen „quasi-science of empirically proving rather common sensical observations" darstellen und daher fordert, dass anstelle der Untersuchung der unmittelbaren Reaktionen auf Selektionsverfahren die Testung von Modellen des Bewerberverhaltens durchzuführen sind.

Der kognitive Pretest und die Reliabilitäts- und Faktorenanalysen haben zudem gezeigt, dass die Akzept!-Skalen für den Einsatz in den Rekrutierungszentren nur bedingt geeignet sind. Man müsste sowohl auf inhaltlicher als auch auf konzeptioneller Ebene eine Überarbeitung vornehmen, indem man zum Beispiel die Satzstruktur vereinfacht und Aussagen, welche in Verbindung mit der Antwortskala zu doppelten Verneinungen führen, überarbeitet. Ein weiteres Problem stellt der starke Zusammenhang der Skalen Messqualität und Augenscheinvalidität dar. Es scheint so zu sein, dass die Stellungspflichtigen diese beiden Konzepte nicht trennen können: Ob ein Test zuverlässig misst oder ob die Aussagen die Alltagsrealität widerspiegeln, sind für sie offenbar Aspekte eines übergeordneten Konstruktes. Auf der Grundlage des noch zu erstellenden Modells müsste somit auch der Akzept!-Fragebogen entsprechend angepasst werden.

7.5 Zusammenhang der Ergebnisse im Leadership-Fragebogen mit anderen Persönlichkeitsmerkmalen und der Intelligenz

Bei nach der traditionellen Methode entwickelten SJT stellt sich die Frage, welche Fähigkeiten des Bewerbers damit gemessen werden. Wie in Kapitel 2.3 aufgeführt, korrelieren die meisten SJT deutlich höher mit Intelligenz (z. B. McDaniel, Morgeson, Finnegan, Campion & Braverman, 2001) als mit Persönlichkeitsdimensionen (z. B. McDaniel, Hartman, Whetzel & Grubb, 2007). Für diese Befunde lassen sich auch Erklärungen finden, beziehen sich SJT doch auf konkrete Arbeitssituationen, welche langjährige Job-Inhaber als erfolgsrelevant bezeichnen. Dies kann auch der Grund sein, weshalb die Ergebnisse aus SJT mit der kognitiven Leistungsfähigkeit korrelieren und es nicht klar ist, was mit SJT überhaupt erfasst wird: Ist es ein Bild der Persönlichkeit des Bearbeiters, indem dieser beschreibt, wie er sich in solchen Situationen verhält oder ist es seine Fähigkeit, sich in die Situationen hineinzuversetzen und auf Grund seines Erfahrungsschatzes zu antizipieren, welches wohl das erfolgversprechendste Verhalten in dieser Situation ist? McDaniel et al. (2007) konnten in ihrer Meta-Analyse nachweisen, dass dies zu einem Teil von der Art der Instruktion abhängt: Dabei korrelieren Antworten, welche auf eine Wissensinstruktion (*should do*) gegeben wurden, eher mit Intelligenz und solche auf eine Verhaltensinstruktion (*would do*) eher mit Persönlichkeitsdimensionen.

Bei der Entwicklung des Leadership-Fragebogens setzte ich mir schon zu Beginn das Ziel, mit dem SJT-Format klar definierte Persönlichkeitsdimensionen zu erfassen. Darauf richtete ich den gesamten Konstruktionsprozess aus und mittels der durchgeführten Faktorenanalysen konnte ich die operationalisierten Dimensionen auch bestätigen. Anhand der Datensätze aus den Rekrutierungszentren, welche auch die Angaben zu den Leistungstests und einer kurzen Persönlichkeitsskala umfassen, werde ich überprüfen, ob die Dimensionen des Leadership-Fragebogens höher mit Persönlichkeitsaspekten korrelieren als mit Intelligenz.

Für diese Berechnungen verwende ich den Rekrutierungs-Datensatz aus dem Jahr 2005, welcher die Angaben von 21'478 deutschschweizer Stellungspflichtigen umfasst. In einem ersten Schritt der Datenbereinigung habe ich alle unvollständigen Datensätze entfernt: Bei 2'269 Stellungspflichtigen fehlen die Angaben zum Leadership-Fragebogen, 406 haben den Intelligenztest nicht bearbeitet, 59 den Textverständnistest und 1'228 den Merkfähigkeitstest. Dies führt zu einem Datensatz mit den vollständigen Angaben zu 17'516 Stellungspflichtigen. Anhand der Analyse der in Box-Plots dargestellten Testbearbeitungszeiten,

schliesse ich Stellungspflichtige mit extrem kurzen oder extrem langen Bearbeitungszeiten aus dem Datensatz aus: Einen der beiden Untertests des Intelligenztests haben insgesamt 76 Stellungspflichtige in weniger als 120 Sekunden bearbeitet, beim Textverständnistest benötigten zwei mehr als 28.3 Minuten und neun weniger als 4.3 Minuten, beim Leadership-Fragebogen drei mehr als 45 Minuten und 362 weniger als 7.5 Minuten und beim Merkfähigkeitstest zwei mehr als 50 Minuten und 22 weniger als 15 Minuten. Dies führt zu einem definitiven Datensatz, welcher 17'040 Stellungspflichtige umfasst. Insgesamt habe ich somit 21.7% Cases ausgeschlossen, 2.7% auf Grund der Bearbeitungsdauer. Das durchschnittliche Alter in der Stichprobe beträgt 19.66 Jahre (*SD* = .93; Range = 17 – 26) und 0.5% der Stellungspflichtigen sind weiblich.

Für die Berechnung der Korrelationen mit der likert-skalierten Version des Leadership-Fragebogens stütze ich mich auf den in Kapitel 7.2 verwendeten Datensatz mit *N* = 1'017 Stellungspflichtigen. Um auch diejenigen Stellungspflichtigen aus dem Datensatz auszuschliessen, welche in einem der Leistungstests Anzeichen für eine unseriöse Bearbeitung zeigen, führe ich eine weitere Datenbereinigung durch, indem ich die Bearbeitungsdauer der Testverfahren als Box-Plots ausgeben lasse. Auf dieser Grundlage schliesse ich vier Stellungspflichtige aus, welche je in einem der Testverfahren eine sehr kurze Bearbeitungszeit aufweisen. Zudem schliesse ich noch 83 Stellungspflichtige aus, welche den Merkfähigkeitstest nicht bearbeitet haben. Dies führt zu einem Datensatz, welcher die Angaben von 930 Stellungspflichtigen umfasst.

Tabelle 7.35

Korrelationen der Scores der Forced-Choice- und der likert-skalierten Version des Leadership-Fragebogens mit den Scores aus den in der Rekrutierung eingesetzten Leistungstests

Dimension	Durchsetzungsfähigkeit		Kontaktfähigkeit		Verantwortungsbewusstsein	
	F-C	likert	F-C	likert	F-C	likert
Textverständnistest	-.02	.02	.01	.04	-.00	.05
Intelligenztest 95, total	-.01	-.07	-.09	-.05	-.08	-.02
Intelligenztest 95, figural	-.02	-.07	-.05	-.01	-.05	-.02
Intelligenztest 95, verbal	-.00	-.03	-.09	-.08	-.08	-.01
Merkfähigkeitstest, total	-.05	-.01	.02	.02	.02	.07
Merkfähigkeitstest, auditiv	-.04	.00	.01	-.02	.02	.04
Merkfähigkeitstest, visuell	-.04	-.02	.03	.05	.02	.08
Intelligenz & Merkfähigkeit	-.03	-.05	-.04	-.03	-.04	.02

Anmerkung. *N* = 17'040 resp. *N* = 930. F-C = Forced-Choice. (Bonferroni-korrigiertes Signifikanzniveau $p < .002$. Korrelationen sind ab .02 resp. .07 signifikant.)

In Tabelle 7.35 habe ich die Korrelationen zwischen den drei in den Rekrutierungszentren eingesetzten Leistungstests und den beiden Versionen des Leadership-Fragebogens aufgeführt. Es ist ersichtlich, dass die Korrelationen bei beiden Leadership-Versionen in etwa gleich hoch ausfallen und dass keine Korrelation über .09 liegt. Dies belegt, dass es gelungen ist, einen SJT zu entwickeln, welcher definitiv von Intelligenzleistungen unabhängige Konstrukte erfasst. In Tabelle 7.36 habe ich zusätzlich die Korrelationen zwischen den Leistungstests und den Skalen eines in den Rekrutierungszentren eingesetzten Persönlichkeits-Fragebogens aufgeführt. Auch hier zeigt sich, dass – wie erwartet – die Korrelationen sehr tief ausfallen. Eine Ausnahme bildet der Merkfähigkeitstest, dessen Resultate einen schwachen Zusammenhang zur selbstbeschriebenen Belastbarkeit aufweisen.

Tabelle 7.36
Korrelationen der Scores von sechs Persönlichkeitsdimensionen mit den Scores aus den Leistungstests

Dimension	LM	BL	EX	GW	EF	TF
Textverständnistest	.07	.14	.04	.04	.13	.07
Intelligenztest 95, total	.01	.08	-.06	-.03	.09	-.05
Intelligenztest 95, figural	.02	.05	-.04	.01	.06	-.03
Intelligenztest 95, verbal	-.01	.08	-.06	-.05	.08	-.05
Merkfähigkeitstest, total	.11	.17	.02	.09	.17	.04
Merkfähigkeitstest, auditiv	.09	.15	.01	.07	.15	.03
Merkfähigkeitstest, visuell	.11	.16	.03	.09	.15	.04
Intelligenz & Merkfähigkeit	.06	.14	-.03	.03	.14	-.01

Anmerkung. N = 16'994. LM = Leistungsmotivation; BL = Belastbarkeit; EX = Extraversion; GW = Gewissenhaftigkeit; EF = Entgegenkommen/Friedfertigkeit; TF = Teamfähigkeit. (Bonferroni-korrigiertes Signifikanzniveau $p < .001$. Korrelationen sind ab .02 signifikant.)

In der nachfolgend dargestellten Tabelle 7.37 sind die Korrelationen zwischen den beiden Versionen des Leadership-Fragebogens und den sechs Dimensionen des Persönlichkeits-Fragebogens aufgeführt. Vor allem bei den Dimensionen Kontaktfähigkeit und Verantwortungsbewusstsein ergeben sich mittlere bis hohe Korrelationen zwischen r = .30 und .73. Die Dimension Durchsetzungsfähigkeit korreliert nur – negativ – mit der Dimension Entgegenkommen/Friedfertigkeit, was inhaltlich durchaus Sinn macht. Auch bei der Kontaktfähigkeit tritt die höchste Korrelation mit der korrespondierenden Persönlichkeitseigenschaft Extraversion auf. Verantwortungsbewusstsein korreliert hoch mit Leistungsmotivation, Belastbarkeit und Gewissenhaftigkeit. Offenbar scheint hier der Zusammenhang

durch die Art und Weise der Arbeitsausführung bedingt zu sein: Wer verantwortungsbewusst ist, ist dies nicht nur in seinem Verhalten anderen Menschen gegenüber, sondern auch gegenüber den auszuführenden Tätigkeiten, welche er gewissenhaft, motiviert und mit hohem Einsatz erledigt.

Tabelle 7.37

Korrelationen der Scores der Forced-Choice- und der likert-skalierten Version des Leadership-Fragebogens mit den Scores aus den in der Rekrutierung eingesetzten Persönlichkeits-Fragebogen

Dimension	Durchsetzungs-fähigkeit		Kontaktfähigkeit		Verantwortungs-bewusstsein	
	F-C	likert	F-C	likert	F-C	likert
Leistungsmotivation	-.09	-.04	.52	.56	**.49**	**.56**
Belastbarkeit	-.05	-.01	.50	.54	**.45**	**.52**
Extraversion	-.01	-.02	**.68**	**.73**	.39	**.46**
Gewissenhaftigkeit	-.08	-.03	.41	.48	**.45**	**.54**
Entgegenkommen/Friedfert.	**-.25**	**-.25**	.33	.36	.30	.38
Teamfähigkeit	-.10	-.06	.49	.52	.35	.37

Anmerkung. N = 17'040 (F-C = Forced-Choice) resp. N = 930 (likert). (Bonferroni-korrigiertes Signifikanzniveau $p < .003$. Korrelationen sind ab .02 resp. .20 signifikant.)

Anhand weiterführender Berechnungen will ich die Korrelationsstruktur noch näher untersuchen. Auffallend bei der in Tabelle 7.37 dargestellten Korrelationsmatrix sind die relativ homogen ausfallenden, mittleren bis hohen Korrelationen zwischen den Persönlichkeitsdimensionen und den Dimensionen Kontaktfähigkeit und Verantwortungsbewusstsein einerseits und die davon deutlich abweichende Korrelationsstruktur bei der Dimension Durchsetzungsfähigkeit. Einen Erklärungsansatz für die homogenen Korrelationen liefert die in Tabelle 7.38 dargestellte Korrelationsmatrix der sechs Skalen des Persönlichkeits-Fragebogens: Alle Skalen korrelieren zwischen r = .35 und .87 miteinander, obwohl sie zum Teil – zumindest theoretisch – unabhängige Konstrukte erfassen, wie zum Beispiel Leistungsmotivation und Entgegenkommen/Friedfertigkeit. Da diese Skalen schon mittel bis hoch miteinander korrelieren, müssen deren Korrelationen mit anderen Konstrukten relativ homogen ausfallen. Der Grund dafür könnte bei bewussten Selbstdarstellungstendenzen (Antwortverzerrungen) liegen, welche die Stellungspflichtigen beim Bearbeiten der Fragebogen einsetzen. So zeigen die Ergebnisse aus Labor- und Feldstudien, dass die bewusste Antwortverzerrung – also das Faking – zu einer deutlichen Erhöhung der Korrelationen zwischen den erfassten Persönlichkeitsdimensionen führt (z. B. Douglas, McDaniel & Snell,

1996; Schmit & Ryan, 1993). Einen Hinweis dafür liefern in meinen Daten die Korrelationen der Skalen des Leadership- und des Persönlichkeits-Fragebogens mit der Kurzsskala Führungsmotivation und dem Einzelitem „Ich kann mir vorstellen, in naher Zukunft eine Kaderposition im Militär inne zu haben", welche ich in den Tabellen 7.39 und 7.40 aufgeführt habe. Die Korrelationen liegen zwischen r = .29 und .51 mit den Ausnahmen Entgegenkommen/Friedfertigkeit (r = .19 / .21) und Durchsetzungsfähigkeit (r = -.06 bis .01).

Tabelle 7.38

Korrelationsmatrix der sechs Skalen des Persönlichkeits-Fragebogens

Dimension	LM	BL	EX	GW	EF	TF
Leistungsmotivation	*.92*	.86	.63	.77	.55	.51
Belastbarkeit	.87	*.88*	.64	.75	.53	.47
Extraversion	.61	.64	*.87*	.51	.37	.61
Gewissenhaftigkeit	.76	.74	.45	*.84*	.48	.40
Entgegenkommen/Friedfert.	.54	.55	.35	.50	*.82*	.42
Teamfähigkeit	.53	.53	.59	.40	.43	*.90*

Anmerkung. N = 16'994 unterhalb der Diagonale; N = 930 oberhalb der Diagonale. Alle Korrelationen sind signifikant. In der Diagonale sind die Skalenreliabilitäten aufgeführt.

Tabelle 7.39

Korrelationen der Scores der Forced-Choice- und der likert-skalierten Version des Leadership-Fragebogens mit der Führungsmotivation und der Bereitschaft, in der Armee eine Kaderposition zu übernehmen

Dimension	Durchsetzungs-fähigkeit		Kontaktfähigkeit		Verantwortungs-bewusstsein	
	F-C	likert	F-C	likert	F-C	likert
Führungsmotivationsskala	-.04	.01	.46	.44	.45	.43
Weitermachen im Militär	-.06	-.05	.42	.43	.42	.41

Anmerkung. N = 16'990 (F-C = Forced-Choice) resp. N = 930 (likert). (Bonferroni-korrigiertes Signifikanzniveau p < .008. Korrelationen sind ab .02 resp. .20 signifikant.)

Tabelle 7.40

Korrelationen der Scores des Persönlichkeits-Fragebogens mit der Führungsmotivation und der Bereitschaft, in der Armee eine Kaderposition zu übernehmen

Dimension	LM	BL	EX	GW	EF	TF
Führungsmotivationsskala	.51	.41	.40	.40	.19	.34
Weitermachen in Militär	.46	.40	.37	.38	.21	.29

Anmerkung. N = 16'994. Alle Korrelationen sind signifikant.

In Tabelle 7.41 habe ich die Partialkorrelationen zwischen den Skalen des Leadership-Fragebogens und denjenigen des Persönlichkeits-Fragebogens unter Kontrolle der Führungsmotivation aufgeführt. Es zeigt sich wie erwartet, dass die Korrelationen leicht sinken, dass sich also die Führungsmotivation auf das Antwortverhalten der Stellungspflichtigen auswirkt.

Tabelle 7.41

Vergleich der Korrelationen mit den Partialkorrelationen der Scores der Forced-Choice-Version des Leadership-Fragebogens mit den Scores des Persönlichkeits-Fragebogen unter Kontrolle der Führungsmotivation

Dimension	Durchsetzungs-fähigkeit		Kontaktfähigkeit		Verantwortungs-bewusstsein	
	F-C	F-C part	F-C	F-C part	F-C	F-C part
Leistungsmotivation	-.09	-.08	.52	.38	**.49**	**.34**
Belastbarkeit	-.05	-.04	.50	.39	**.45**	**.32**
Extraversion	-.01	.01	**.68**	**.61**	.39	.25
Gewissenhaftigkeit	-.08	-.07	.41	.28	**.45**	**.33**
Entgegenkommen/Friedfert.	**-.25**	**-.25**	.33	.28	.30	.25
Teamfähigkeit	-.10	-.09	.49	.40	.35	.23

Anmerkung. N = 17'040 F-C = Forced-Choice. (Bonferroni-korrigiertes Signifikanzniveau $p < .003$. Korrelationen sind ab .02 signifikant.)

Abschliessend stelle ich die Korrelationen des Leadership-Fragebogens mit zwei weiteren Persönlichkeitsskalen dar. Für eine Validierungsstudie haben zwei Studentinnen (Imper & Maier, 2003) einen kurzen Persönlichkeits-Fragebogen zu den drei Leadership-Dimensionen entwickelt, welcher insgesamt 24 sechsstufig likert-skalierte Items umfasst. Nachfolgend führe ich beispielhaft je ein Item zu den drei Dimensionen auf: „Bei Auseinandersetzungen gewinne ich andere leicht für meine Position", „Ich kann besser auf Menschen zugehen als viele andere.", „Ich übernehme gerne verantwortungsvolle Aufgaben." Die Reliabilitäten dieser drei Skalen betragen α = .71, .88 und .83, wobei auch hier diejenige zur Durchsetzungsfähigkeit mit Abstand den tiefsten Wert aufweist. An dieser Studie haben 442 Kaderanwärter teilgenommen, welche die beiden Testverfahren im Rahmen der Kaderbeurteilung Stufe II für zukünftige Zugführer, Feldweibel und Fouriere ausgefüllt haben. Der Tabelle 7.42 lässt sich entnehmen, dass die Skalen jeweils am höchsten mit derjenigen korrelieren, welche dasselbe Merkmal erfasst. Deutlich zeigt sich jedoch wieder der Unterschied zwischen der Skala Durchsetzungsfähigkeit und den beiden anderen Skalen.

Tabelle 7.42

Doppelt minderungskorrigierte Korrelationen der Scores der Forced-Choice-Version des Leadership-Fragebogens mit den Scores aus dem Vergleichsfragebogen

Leadership-Fragebogen		Durchsetzungs-fähigkeit	Kontaktfähig-keit	Verantwortungs-bewusstsein
Vergleichsfragebogen	α	*.52*	*.79*	*.60*
Durchsetzungsfähigkeit	*.71*	**.26**	.45	.29
Kontaktfähigkeit	*.88*	.07	**.76**	.34
Verantwortungsbewusstsein	*.83*	-.08	.56	**.67**

Anmerkung. N = 442. (Bonferroni-korrigiertes Signifikanzniveau $p < .006$. Korrelationen sind ab .20 signifikant.) Kursiv sind die Reliabilitäten (Cronbach Alpha) aufgeführt.

In Tabelle 7.43 habe ich die doppelt minderungskorrigierten Korrelationen der beiden Leadership-Fragebogen-Versionen mit den Skalen des NEO-PI-R (Ostendorf & Angleitner, 2004) aufgeführt. (Die Matrizen mit den Korrelationen zwischen den Facetten des NEO-PI-R und den beiden Versionen des Leadership-Fragebogens habe ich in Anhang 7.25a und 7.25b abgebildet.) Als Probanden für diese Datenerhebung dienten uns Psychologie-Studierende, welche für die Teilnahme an der Studie zwei Versuchspersonenstunden (zu erbringende Studienleistung) gutgeschrieben bekamen. Ihr durchschnittliches Alter beträgt 22.65 Jahre (*SD* = 3.69), 51% der Probanden sind männlich.

In Tabelle 7.43 ist ersichtlich, dass sich die Werte bei den beiden Leadership-Fragebogen-Versionen kaum unterscheiden, was bedeutet, dass das Antwortformat keinen Einfluss auf die Konstruktvalidität des Fragebogens hat.

Tabelle 7.43

Doppelt minderungskorrigierte Korrelationen der Scores der beiden Versionen des Leadership-Fragebogens mit den Scores des NEO-PI-R

Dimension		Durchsetzungs-fähigkeit		Kontaktfähigkeit		Verantwortungs-bewusstsein	
		F-C	likert	F-C	likert	F-C	likert
	α	*.63*	*.83*	*.83*	*.91*	*.54*	*.79*
Neurotizismus	*.94*	-0.14	-0.25	**-0.39**	**-0.45**	**-0.33**	**-0.40**
Extraversion	*.90*	0.03	0.12	**0.72**	**0.75**	0.26	**0.58**
Offenheit	*.87*	-0.05	-0.00	0.03	0.00	0.01	0.14
Verträglichkeit	*.85*	**-0.48**	**-0.45**	-0.11	-0.17	-0.03	0.10
Gewissenhaftigkeit	*.93*	**0.33**	**0.32**	-0.12	-0.10	0.26	0.24

Anmerkung. N = 100. F-C = Forced-Choice. Es sind die doppelt minderungskorrigierten Werte aufgeführt. (Bonferroni-korrigiertes Signifikanzniveau $p < .003$. Korrelationen sind ab .30 signifikant.) Kursiv sind die Reliabilitäten (Cronbach Alpha) aufgeführt.

Die Skala Durchsetzungsfähigkeit korreliert mit den Big Five-Dimensionen Verträglichkeit (negativ) und Gewissenhaftigkeit, was sich inhaltlich gut erklären lässt: Wer häufig seine Interessen gegenüber anderen Personen durchsetzt, womöglich noch mit einer gewissen Rücksichtslosigkeit, lässt sich mit Adjektiven, welche für eine niedrige Ausprägung in Verträglichkeit stehen beschreiben: dickköpfig, ichbezogen, manipulierend, rechthaberisch, unnachgiebig (Ostendorf & Angleitner, 2004, S. 44). Gewissenhafte Menschen lassen sich als beharrlich, ehrgeizig, motiviert, unbeirrbar, willensstark und zielstrebig beschreiben (Ostendorf & Angleitner, 2004, S. 46), alles Eigenschaften, welche erforderlich sind, wenn man seine Meinung gegenüber anderen durchsetzen will.

Auffällig ist die hohe Korrelation zwischen der Kontaktfähigkeit und der Big Five-Dimension Extraversion. Dies ist inhaltlich auch gerechtfertigt, erfassen diese beiden Skalen doch genau dasselbe Konstrukt. Nach Ostendorf und Angleitner (2004, S. 40), lassen sich extravertierte Menschen mit den Adjektiven aufgeweckt, freundschaftlich, geschwätzig, gesellig, gesprächig, kontaktfähig oder personenorientiert beschreiben. Die negative Korrelation zur Dimension Neurotizismus ist schon auf Ebene des NEO-PI-R gegeben – Extraversion und Neurotizismus korrelieren mit $r = -.27$ –, was durchaus nachvollziehbar ist, da ängstliche, reizbare, depressive oder sozial befangene Personen den Kontakt zu Mitmenschen eher meiden.

Weniger klar sind die Zusammenhänge bei der Skala Verantwortungsbewusstsein, nicht zuletzt auch deshalb, weil diese als solche nicht in den Big Five abgebildet ist. Da es sich jedoch um eine Eigenschaft im Umgang mit Menschen handelt, lässt sich die Korrelation der likert-skalierten Version des Leadership-Fragebogens mit Extraversion erklären. Die negative Korrelation zu Neurotizismus lässt sich vielleicht damit begründen, dass neurotizistisch veranlagte Menschen stark mit sich selbst und mit ihren eigenen Problemen beschäftigt sind, so dass sie nicht noch den Willen und die Kapazität haben, sich um andere Menschen zu kümmern.

Somit deckt der Leadership-Fragebogen alle Big Five-Dimensionen ab mit Ausnahme der Offenheit, welche im NEO-PI-R mit den Facetten Offenheit für Fantasie, für Ästhetik, für Gefühle, für Handlungen, für Ideen und für Werte- und Normensysteme operationalisiert ist und somit deutlich andere Konstrukte erfasst, als der Leadership-Fragebogen. Auf Facettenebene zeigen sich nur bei Offenheit für Handlungen (experimentierfreudig, flexibel, sucht neue Aktivitäten, bevorzugt Abwechslung; Ostendorf & Angleitner, 2004, S. 36) Zusammenhänge zu Kontaktfähigkeit und Verantwortungsbewusstsein.

Welche Schlussfolgerungen lassen sich aus den in diesem Kapitel präsentierten Korrelationen ziehen?

1. Der Leadership-Fragebogen erfasst keine Intelligenz-Aspekte.
2. Die Dimensionen des Leadership-Fragebogens weisen Zusammenhänge mit verwandten Persönlichkeitseigenschaften auf.
3. Zwei der drei Skalen des Leadership-Fragebogens korrelieren mit Führungsmotivation.

Damit liegt eine – zumindest partielle – Bestätigung der Erreichung eines der Ziele dieser Testkonstruktion vor: Der als Situational Judgment Test konzipierte Leadership-Fragebogen erfasst a priori definierte Persönlichkeitsdimensionen. In Abgrenzung zu den nach der üblichen Vorgehensweise entwickelten SJTs sind die Scores aus dem Leadership-Fragebogen zudem unabhängig von Intelligenzmassen.

Die relativ homogenen Korrelationen zu den Persönlichkeitsdimensionen und die hohen Korrelationen zur Führungsmotivation lassen sich höchstwahrscheinlich auf die bewussten Selbstdarstellungstendenzen der Stellungspflichtigen zurückführen. Aus diesem Grund ist noch eine weitere Validierungsstudie mit einem validierten Persönlichkeits-Fragebogen in einem Labor-Setting durchzuführen, in welchem die Probanden dazu angehalten werden, die Fragebogen ehrlich auszufüllen. Zudem ist das Antwortverhalten bei der Skala Durchsetzungsfähigkeit im Vergleich zu den beiden anderen Skalen zu untersuchen. Dabei soll eine Antwort auf die Frage gefunden werden, ob die abweichenden Korrelationsmuster auf konstruktionsbedingte Mängel in der Skala Durchsetzungsfähigkeit zurückzuführen sind oder ob diese Skala die Besonderheit aufweist, dass sich die Testbearbeiter nicht im Klaren darüber sind, welche der in den Items geschilderten Verhaltensweisen als die effektivsten im Bezug auf eine Führungstätigkeit sind. Würde dies zutreffen, so wäre dies ein Hinweis auf die Art des Zusammenhanges zwischen der Intransparenz einer Skala und deren Validität.

7.6 Literaturverzeichnis

Anderson, N. (2003). Applicant and recruiter reactions to new technology in selection: A critical review and agenda for future research. *International Journal of Selection and Assessment, 11*, 121–136.

Barrick, M. R., & Mount, M. K. (2003). Impact of meta-analysis methods on understanding personality–performance relations. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 197–222). Mahwah, NJ: Erlbaum.

Boss, P. (1999). Welche Faktoren beeinflussen die Verständlichkeit von Dokumenten? In P. Notter, E.-M. Bonerad & F. Stoll (Hrsg.), *Lesen – eine Selbstverständlichkeit? Schweizer Bericht zum "International Adult Literacy Survey"* (S. 235–250). Zürich: Rüegger.

Boss, P. (2005). Assessment in der Arbeitswelt – Kriterien für eine bewerberzentrierte Personalauswahl. In M. Rehbinder (Hrsg.), *Psychologische Aspekte im Recht der Personalführung* (S. 21–45). Bern: Stämpfli.

Boss, P. & Baumann, R. (2003). Psychologische Testverfahren beim Rekrutierungsprozess der Armee. *HR-Today, 7–8*, 22–23.

Buss, D. M., & Craik, K. H. (1983). The Act Frequency Approach to personality. *Psychological Review, 90,* 105–126.

Buss, D. M., & Craik, K. H. (1986). The Act Frequency Approach and the construction of personality. In A. Angleitner, A. Furnham, & G. Van Heck (Eds.), *Personality psychology in Europe. Volume 2. Current trends and controversies* (pp. 141–156). New York, NY: Academic Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Gleser; G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Cropanzano, R. (1994). The justice dilemma in employee selection: Some reflections on the trade-offs between fairness and validity. *The Industrial–Organizational Psychologist, 31*, 90–93.

Cropanzano, R., & Wright, T. A. (2003). Procedural justice and organizational staffing: A tale of two paradigms. *Human Resource Management Review, 13*, 7–39.

Deiss, E., Emerson, A., Imper, A. & Maier, V. (2002). *Überprüfung und Weiterentwicklung des Leadership-Fragebogens zur Erfassung von Führungspotenzial.* Unveröff. Bericht, Universität Zürich.

Dineen, B. R., Noe, R. A., & Wang, C. (2004). Perceived fairness of web-based applicant screening procedures: Weighing the rules of justice and the role of individual differences. *Human Resource Management, 43*, 127–145.

Douglas, E. F., McDaniel, M. A., & Snell, A. F. (1996, August). The validity of non-cognitive measures decays when applicants fake. *Academy of Management Proceedings*, 127–131.

Eberle, W. & Hartwich, E. (1995). *Brennpunkt Führungspotential. Persönlichkeitseinschätzung als unternehmerische Aufgabe.* Frankfurt am Main: Frankfurter Allgemeine Zeitung.

Evers, A. (2001). The revised Dutch rating system for test quality. *International Journal of Testing, 1*, 155–182.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *The Academy of Management Review, 18*, 694–734.

Gloor, A. (2007). Die Verheiratung des Big-Five-Konzeptes mit dem Wertequadrat-Modell – ein Entwurf. In F. Westermann (Hrsg.), *Entwicklungsquadrat. Theoretische Fundierung und praktische Anwendungen* (S. 31–44). Göttingen: Hogrefe.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26*, 499–510.

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*, 265–275.

Haitovsky, Y. (1969). Multicollinearity in regression analysis: A comment. *Review of Economics and Statistics, 51*, 486–489.

Hausknecht, J. P., Day, D. V., & Thomas, S. C. (2004). Applicant reactions to selection procedures: An updated model and meta-analysis. *Personnel Psychology, 57*, 639–683.

Helwig, P. (1948). Das Wertequadrat. *Psyche, 2*, 121–127.

Hough, L. M., & Oswald, F. L. (2008). Personality testing and industrial-organizational psychology: Reflections, progress, and prospects. *Industrial and*

Organizational Psychology: Perspectives on Science and Practice, 1, 272–290.

Imper, A. & Maier, V. (2003). *Validierung eines Fragebogens zur Erfassung von Führungspotenzial bei Stellungspflichtigen.* Unveröff. Bericht, Universität Zürich.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*, 31–36.

Kersting, M. (1998). Differenzielle Aspekte der sozialen Akzeptanz von Intelligenztests und Problemlöseszenarien als Personalauswahlverfahren. *Zeitschrift für Arbeits- und Organisationspsychologie, 42*, 61–75.

Kersting, M. (Juni 2006). *Akzeptanz in der psychologischen Diagnostik.* Unterlagen zum Vortrag am Psychologischen Institut der Universität Zürich, Fachrichtung Persönlichkeitspsychologie und Diagnostik, Zürich.

Kersting, M. (2008). Zur Akzeptanz von Intelligenz- und Leistungstests. *Report Psychologie, 33*, 420–433.

Klare, G. R. (1963). *The measurement of readability.* Ames: The Iowa State University Press.

Kommando Rekrutierung (2009). *Rekrutierungsbericht 2007 / 2008.* Bern: Kommando Rekrutierung.

Krüger, C. & Amelang, M. (1995). Bereitschaft zu riskantem Verhalten als Trait-Konstrukt und Test-Konzept: Zur Entwicklung eines Fragebogens auf der Basis des Handlungs-Häufigkeits-Ansatzes. *Diagnostica, 41*, 35–52.

Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Psychologie Verlags Union.

Lindley, P., Bartram, D., & Kennedy, N. (2008). *EFPA review model for the description and evaluation of psychological tests. Test review form and notes for reviewers* (Version 3.42). European Federation of Psychologists' Associations. Heruntergeladen am 1. Juli 2010 von www.efpa.eu/download/9044bd41c7953b956876e06c797f8c9f

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51*, 493–504.

Madigan, J., & Macan, T. H. (2005). Improving applicant reactions by altering test administration. *Applied H.R.M. Research, 10*, 73–87.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L., III. (2007). Situational judgment tests, response instructions, and validity: A

meta-analysis. *Personnel Psychology, 60,* 63–91.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Predicting job performance using situational judgment tests: A clarification of the literature. *Journal of Applied Psychology, 86,* 730–740.

Ostendorf, F. & Angleitner, A. (2004). *NEO-PI-R. NEO-Persönlichkeitsinventar nach Costa und McCrae. Revidierte Fassung.* Göttingen: Hogrefe.

Prüfer, P. & Rexroth, M. (2000). Zwei-Phasen-Pretesting. In P. P. Mohler & P. Lüttinger (Hrsg.), *Querschnitt. Festschrift für Max Kaase* (S. 203–219). Mannheim: ZUMA.

Rockwell, R. C. (1975). Assessment of multicollinearity: The Haitovsky test of the determinant. *Sociological Methods and Research, 3,* 308–320.

Rost, J. (1996). *Lehrbuch Testtheorie, Testkonstruktion.* Bern: Huber.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review, 16*, 155–180.

Rothstein, M. G., & Jelly, R. B. (2003). The challenge of aggregating studies of personality. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 223–262). Mahwah, NJ: Erlbaum.

Rückert, J. (1993). *Psychometrische Grundlagen der Diagnostik.* Göttingen: Hogrefe.

Rudolf, M. & Müller, J. (2004). *Multivariate Verfahren. Eine praxisorientierte Einführung mit Anwendungsbeispielen in SPSS.* Göttingen: Hogrefe.

Ryan, A. M., Greguras, G. J., & Ployhart, R. E. (1996). Perceived job relatedness of physical ability testing for firefighters: Exploring variations in reaction. *Human Performance, 9*, 219–240.

Schendera, C. F. G. (2008). *Regressionsanalyse mit SPSS.* München: Oldenbourg.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262–274.

Schmit, M. J., & Ryan, A. M. (1993). The Big Five in personnel selection: Factor structure in applicant and nonapplicant populations. *Journal of Applied Psychology*, 78, 966–974.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston, MA: Allyn and Bacon.

Tauber, M., Stoll, F. & Drewek, R. (1980). Erfassen Lesbarkeitsformeln und Textbeurteilung verschiedene Dimensionen der Textverständlichkeit? *Zeitschrift für experimentelle und angewandte Psychologie, 1*, 135–146.

Van den Bos, K., Bruins, J., Wilke, H. A. M., & Dronkert, E. (1999). Sometimes unfair procedures have nice aspects: On the psychology of the Fair Process Effect. *Journal of Personality and Social Psychology, 77*, 324–366.

Warr, P., Bartram, D., & Martin, T. (2005). Personality and sales performance: Situational variation and interactions between traits. *International Journal of Selection and Assessment, 13*, 87–91.

Anhang 7.1 Vergleich der anhand der Datensätze 2008 (*N* = 19'801) und 2003 (*N* = 7'871) berechneten Itemkennwerte des Leadership-Fragebogens

Durchsetzungsfähigkeit	Datensatz 2008				Datensatz 2003	
Item	*M*	*SD*	r_{it}	Faktor-ladung	r_{it}	Faktor-ladung
Lohnerhöhung	2.58	0.86	.22	.41	.20	.39
Fahrer	2.55	0.80	.28	.47	.25	.48
Unterbruch	2.01	0.81	.26	.47	.24	.45
Aufräumen	2.70	0.71	.29	.51	.26	.47
Waschküche	1.91	0.61	.27	.48	.25	.47
Geschirr	2.81	0.79	.24	.45	.23	.43
Musik	2.36	0.77	.24	.43	.19	.39
Probleme	2.00	0.99	.24	.41	.21	.42
Auswärts	2.56	0.89	.25	.42	.23	.44
Arbeit	2.66	0.89	.26	.45	.23	.44
Cronbach Alpha			.57		.53	

Kontaktfähigkeit	Datensatz 2008				Datensatz 2003	
Item	*M*	*SD*	r_{it}	Faktor-ladung	r_{it}	Faktor-ladung
Nachbarn	2.88	0.72	.55	.56	.54	.55
Zugfahrt	2.61	0.74	.62	.69	.60	.70
Kurs	2.53	0.74	.57	.61	.54	.58
Schultag	2.77	0.95	.50	.59	.49	.57
Flugzeug	2.67	0.90	.58	.65	.57	.65
Barmann	2.33	0.89	.44	.60	.45	.60
Begleitung	2.46	0.94	.55	.65	.51	.63
Umzug	2.81	0.79	.51	.53	.48	.49
Geburtstag	2.51	0.81	.51	.61	.50	.61
Fitness	2.61	0.87	.52	.63	.53	.63
Cronbach Alpha			.84		.83	

Verantwortungsbewusstsein	Datensatz 2008				Datensatz 2003	
Item	*M*	*SD*	r_{it}	Faktor-ladung	r_{it}	Faktor-ladung
Schanze	2.75	0.88	.31	.50	.31	.50
Beratungsstelle	2.64	0.66	.36	.47	.34	.46
Silvester	2.97	0.97	.41	.52	.40	.51
Subventionen	2.18	0.73	.33	.40	.35	.40
Nachhilfestunden	2.47	0.75	.37	.45	.40	.51
Kind	2.64	0.85	.38	.48	.36	.46
Wohnung	2.63	0.80	.29	.49	.33	.48
Bergtour	2.80	0.88	.38	.55	.38	.52
Ampel	2.92	0.93	.43	.55	.43	.54
Autofahren	3.01	0.84	.39	.51	.39	.53
Cronbach Alpha			.70		.71	

Anhang 7.2 Item-Korrelationsmatrix (Datensatz 2008; *N* = 19'801)

	Items der Skala Durchsetzungsfähigkeit												Items der Skala Kontaktfähigkeit												Items der Skala Verantwortungsbewusstsein											
	02	04	09	13	15	19	21	22	24	27	03	06	08	10	14	16	18	25	29	30	01	05	07	11	12	17	20	23	26	28
02	-	.09	.13	.12	.12	.11	.09	.09	.08	.10	.04	.05	.05	.06	.04	.05	.06	.02	.05	.03	.00	.04	.01	.03	.00	.04	.02	.04	.00	.05
04		-	.14	.12	.10	.10	.10	.16	.20	.14	-.07	-.07	-.05	-.03	-.07	-.04	-.05	-.06	-.05	-.06	-.01	-.03	-.07	-.07	-.04	-.03	.04	.01	-.05	-.01
09			-	.14	.15	.12	.13	.11	.10	.09	-.01	.01	.00	.01	-.02	.02	.01	-.02	.02	-.02	.00	.04	.00	-.02	.01	.02	.06	.04	-.01	.04
13				-	.15	.20	.14	.10	.09	.15	-.01	-.01	-.02	.01	-.02	.00	.00	-.02	.01	-.02	.00	.02	-.03	-.01	.02	-.02	.06	.03	-.02	.03
15					-	.15	.14	.09	.10	.13	-.01	.00	.02	.03	.01	.02	.04	-.01	.03	.01	.01	.01	.00	.04	.02	.01	.10	.05	.02	.04
19						-	.09	.05	.09	.12	.07	.05	.06	.08	.06	.03	.06	.07	.06	.04	.07	.07	.06	.05	.05	.06	.14	.09	.09	.11
21							-	.11	.10	.11	-.13	-.11	-.10	-.07	-.10	-.03	-.06	-.12	-.06	-.08	-.07	-.06	-.10	-.06	-.09	-.10	-.04	-.06	-.13	-.04
22								-	.16	.13	-.10	-.10	-.08	-.07	-.10	-.06	-.07	-.10	-.06	-.08	-.09	-.07	-.08	-.09	-.08	-.06	-.02	-.04	-.10	-.07
24									-	.12	-.09	-.10	-.08	-.07	-.10	-.08	-.07	-.08	-.08	-.08	-.03	-.04	-.07	-.08	-.08	-.05	.03	-.03	-.07	-.03
27										-	-.02	-.03	.00	.00	-.02	-.02	.00	-.01	-.01	-.02	.01	.01	-.01	-.01	-.02	-.01	.05	.03	.00	.02
03											-	.42	.40	.34	.40	.26	.33	.41	.30	.34	.16	.23	.27	.24	.27	.23	.16	.20	.27	.23
06												-	.43	.36	.51	.33	.40	.37	.37	.40	.12	.22	.25	.23	.25	.23	.12	.19	.26	.20
08													-	.38	.41	.29	.36	.36	.33	.34	.14	.21	.22	.20	.25	.23	.13	.19	.26	.22
10														-	.35	.27	.34	.31	.31	.30	.12	.18	.18	.19	.19	.18	.11	.14	.20	.18
14															-	.29	.37	.36	.32	.37	.12	.21	.23	.23	.22	.22	.13	.18	.27	.20
16																-	.31	.23	.30	.32	.08	.12	.12	.13	.16	.14	.04	.09	.13	.14
18																	-	.32	.41	.34	.10	.18	.20	.20	.19	.18	.10	.16	.21	.18
25																		-	.32	.33	.14	.18	.24	.21	.23	.20	.15	.19	.26	.22
29																			-	.32	.11	.17	.18	.18	.19	.18	.10	.14	.22	.19
30																				-	.11	.16	.18	.14	.20	.19	.09	.13	.21	.18
01																					-	.16	.17	.13	.17	.15	.16	.18	.19	.19
05																						-	.22	.16	.21	.20	.14	.19	.21	.22
07																							-	.21	.21	.25	.15	.21	.28	.22
11																								-	.18	.20	.17	.18	.19	.16
12																									-	.21	.14	.20	.24	.21
17																										-	.13	.20	.27	.21
20																											-	.19	.18	.17
23																												-	.23	.22
26																													-	.23
28																														-

Anhang 7.3 Überprüfung der Wertequadrate der Skala Durchsetzungsfähigkeit (Datensatz 2008; *N* = 19'801)

	r_{it}		*n*	%	min	max	*M*	*SD*
Lohnerhöhung	.22	WQ 1	2'185	11.03	10	35	21.31	3.77
		WQ 2	6'753	34.10	12	37	23.07	3.08
		WQ 3	8'141	41.11	15	36	24.80	3.17
		WQ 4	2'722	13.75	16	40	26.99	3.93
Fahrer	.28	WQ 1	1'006	5.08	10	36	20.68	4.14
		WQ 2	9'879	49.89	11	37	22.98	3.09
		WQ 3	6'020	30.40	12	38	25.02	3.09
		WQ 4	2'896	14.63	15	40	27.39	3.72
Unterbruch	.26	WQ 1	4'681	23.64	10	34	21.83	3.39
		WQ 2	11'947	60.34	12	38	24.17	3.08
		WQ 3	1'466	7.40	15	39	26.83	3.86
		WQ 4	1'707	8.62	17	40	27.79	3.64
Aufräumen	.29	WQ 1	349	1.76	10	35	21.20	4.96
		WQ 2	7'784	39.31	11	34	22.38	3.18
		WQ 3	9'150	46.21	12	38	24.83	3.09
		WQ 4	2'518	12.72	13	40	27.38	3.84
Waschküche	.27	WQ 1	4'152	20.97	10	34	21.96	3.65
		WQ 2	13'716	69.27	14	38	24.25	3.16
		WQ 3	1'431	7.23	16	39	27.65	3.69
		WQ 4	502	2.54	16	40	28.71	4.41
Geschirr	.24	WQ 1	1'233	6.23	10	34	20.62	4.06
		WQ 2	4'654	23.50	13	34	22.75	3.17
		WQ 3	10'551	53.29	14	37	24.22	3.15
		WQ 4	3'363	16.98	15	40	27.02	3.68
Musik	.24	WQ 1	1'383	6.98	10	35	20.77	3.67
		WQ 2	12'075	60.98	11	35	23.42	3.15
		WQ 3	4'150	20.96	15	38	25.80	3.22
		WQ 4	2'193	11.08	15	40	26.98	4.12
Probleme	.24	WQ 1	8'514	43.00	10	34	22.36	3.21
		WQ 2	4'004	20.22	13	36	24.10	3.07
		WQ 3	6'109	30.85	14	38	25.82	3.13
		WQ 4	1'174	5.93	13	40	28.23	4.19
Auswärts	.25	WQ 1	2'153	10.87	10	36	21.36	3.60
		WQ 2	7'691	38.84	12	36	23.02	3.04
		WQ 3	6'693	33.80	12	38	24.81	3.20
		WQ 4	3'264	16.48	16	40	27.17	3.61
Arbeit	.26	WQ 1	2'213	11.18	10	35	21.26	3.47
		WQ 2	5'715	28.86	11	36	22.50	2.99
		WQ 3	8'526	43.06	12	39	24.94	3.28
		WQ 4	3'347	16.90	16	40	26.75	3.36

Anhang 7.4 Überprüfung der Wertequadrate der Skala Kontaktfähigkeit (Datensatz 2008; *N* = 19'801)

	r_{it}		*n*	%	min	max	*M*	*SD*
Nachbarn	.55	WQ 1	842	4.25	10	35	16.18	5.20
		WQ 2	3'970	20.05	11	35	21.40	4.04
		WQ 3	11'753	59.36	12	38	27.40	3.96
		WQ 4	3'236	16.34	14	40	30.22	3.96
Zugfahrt	.62	WQ 1	1'409	7.12	10	33	17.23	5.04
		WQ 2	6'657	33.62	11	37	23.14	3.91
		WQ 3	10'045	50.73	12	39	28.45	3.53
		WQ 4	1'690	8.53	16	40	32.13	3.33
Kurs	.57	WQ 1	2'422	12.23	10	34	18.91	5.09
		WQ 2	5'016	25.33	11	35	23.56	3.93
		WQ 3	11'832	59.75	13	39	28.53	3.83
		WQ 4	531	2.68	18	40	31.82	4.04
Schultag	.50	WQ 1	2'165	10.93	10	33	18.67	5.09
		WQ 2	5'236	26.44	11	37	24.34	4.26
		WQ 3	7'377	37.26	12	37	26.79	3.95
		WQ 4	5'023	25.37	14	40	30.44	3.66
Flugzeug	.58	WQ 1	1'553	7.84	10	35	17.68	5.25
		WQ 2	7'631	38.54	12	36	23.64	3.96
		WQ 3	6'344	32.04	13	37	28.21	3.43
		WQ 4	4'273	21.58	15	40	30.80	3.48
Barmann	.44	WQ 1	3'881	19.60	10	35	21.75	5.60
		WQ 2	7'347	37.10	11	37	24.45	4.14
		WQ 3	6'736	34.02	13	38	29.61	3.58
		WQ 4	1'837	9.28	13	40	29.89	4.08
Begleitung	.55	WQ 1	3'314	16.74	10	36	20.84	5.42
		WQ 2	6'958	35.14	11	35	23.73	3.74
		WQ 3	6'586	33.26	15	38	29.41	3.25
		WQ 4	2'943	14.86	16	40	30.76	3.51
Umzug	.51	WQ 1	1'772	8.95	10	34	19.03	5.67
		WQ 2	3'118	15.75	11	34	21.99	4.35
		WQ 3	12'080	61.01	13	38	27.25	4.01
		WQ 4	2'831	14.30	15	40	30.71	3.78
Geburtstag	.51	WQ 1	982	4.96	10	34	18.80	6.11
		WQ 2	10'683	53.95	11	37	23.97	4.37
		WQ 3	5'127	25.89	13	38	29.32	3.47
		WQ 4	3'009	15.20	16	40	31.09	3.47
Fitness	.52	WQ 1	1'521	7.68	10	34	18.75	6.01
		WQ 2	8'217	41.50	11	36	23.46	3.89
		WQ 3	6'501	32.83	13	39	29.29	3.49
		WQ 4	3'562	17.99	14	40	29.96	3.78

Anhang 7.5 Überprüfung der Wertequadrate der Skala Verantwortungsbewusstsein (Datensatz 2008; *N* = 19'801)

	r_{it}		*n*	%	min	max	*M*	*SD*
Schanze	.31	WQ 1	1'756	8.87	10	34	21.89	4.95
		WQ 2	5'520	27.88	13	36	25.75	3.81
		WQ 3	8'518	43.02	12	37	27.50	3.60
		WQ 4	4'007	20.24	13	39	29.99	3.58
Beratungsstelle	.36	WQ 1	790	3.99	10	33	19.26	4.71
		WQ 2	6'812	34.40	11	36	25.41	3.95
		WQ 3	10'903	55.06	14	38	28.13	3.57
		WQ 4	1'296	6.55	16	39	30.80	3.70
Silvester	.41	WQ 1	2'316	11.70	10	33	21.68	4.32
		WQ 2	2'701	13.64	11	35	24.32	3.78
		WQ 3	8'081	40.81	14	37	27.36	3.35
		WQ 4	6'703	33.85	14	39	29.53	3.36
Subventionen	.33	WQ 1	2'974	15.02	10	35	23.46	4.89
		WQ 2	11'221	56.67	11	37	26.57	3.78
		WQ 3	4'753	24.00	13	38	29.55	3.26
		WQ 4	853	4.31	15	39	31.16	3.52
Nachhilfestunden	.37	WQ 1	2'042	10.31	10	33	21.44	4.74
		WQ 2	7'447	37.61	12	38	25.96	3.71
		WQ 3	9'189	46.41	14	38	28.80	3.38
		WQ 4	1'123	5.67	15	39	29.61	3.72
Kind	.38	WQ 1	2'639	13.33	10	35	21.62	4.21
		WQ 2	4'130	20.86	12	36	25.65	3.72
		WQ 3	10'750	54.29	14	39	28.32	3.46
		WQ 4	2'282	11.52	15	39	29.57	3.44
Wohnung	.29	WQ 1	1'858	9.38	10	35	22.02	4.80
		WQ 2	5'705	28.81	12	36	25.77	3.85
		WQ 3	10'082	50.92	13	39	28.07	3.65
		WQ 4	2'156	10.89	13	39	29.69	3.99
Bergtour	.38	WQ 1	831	4.20	10	34	20.59	5.05
		WQ 2	7'457	37.66	11	36	24.97	3.87
		WQ 3	6'270	31.67	13	37	27.84	3.37
		WQ 4	5'243	26.48	14	39	29.95	3.34
Ampel	.43	WQ 1	2'830	14.29	10	34	22.26	4.01
		WQ 2	999	5.05	11	35	22.49	4.83
		WQ 3	10'899	55.04	13	39	27.14	3.36
		WQ 4	5'073	25.62	15	39	30.30	3.02
Autofahren	.39	WQ 1	725	3.66	10	35	19.37	4.81
		WQ 2	4'722	23.85	11	36	24.26	3.96
		WQ 3	7'940	40.10	13	37	27.41	3.31
		WQ 4	6'414	32.39	13	39	29.42	3.60

Anhang 7.6 Itemkennwerte der Skalen der likert-skalierten Version des Leadership-Fragebogens (Datensatz 2007; *N* = 1'017)

		Datensatz 2007 4 Wertequadranten pro Item (N = 1'017)			*Datensatz 2007 3 Wertequadranten pro Item (N = 1'017)*			*Datensatz 2008 1 Sub-Item pro Item (N = 19'801)*		
	Item	*M*	*SD*	r_{it}	*M*	*SD*	r_{it}	*M*	*SD*	r_{it}
Durchsetzungsfähigkeit	Lohnerhöhung	10.23	2.09	.21	8.16	1.83	.25	2.58	0.86	.22
	Fahrer	10.64	2.58	.40	8.82	2.15	.41	2.55	0.80	.28
	Unterbruch	10.23	1.97	.39	7.01	1.70	.42	2.01	0.81	.26
	Aufräumen	10.94	2.01	.46	9.00	1.60	.48	2.70	0.71	.29
	Waschküche	9.79	2.06	.43	6.31	1.83	.39	1.91	0.61	.27
	Geschirr	10.93	1.95	.42	8.39	1.72	.46	2.81	0.79	.24
	Musik	11.43	1.89	.43	8.06	1.62	.45	2.36	0.77	.24
	Probleme	9.02	2.01	.30	6.49	1.76	.30	2.00	0.99	.24
	Auswärts	10.14	2.09	.39	7.92	1.75	.40	2.56	0.89	.25
	Arbeit	10.79	2.61	.41	8.23	1.95	.38	2.66	0.89	.26
	Cronbach Alpha			.72			.73			.57
Kontaktfähigkeit	Nachbarn	11.24	2.28	.69	8.83	1.93	.67	2.88	0.72	.55
	Zugfahrt	10.39	2.48	.74	8.23	2.09	.72	2.61	0.74	.62
	Kurs	10.08	2.07	.62	7.60	1.68	.66	2.53	0.74	.57
	Schultag	11.48	2.05	.61	8.58	1.76	.66	2.77	0.95	.50
	Flugzeug	11.61	2.27	.65	8.23	2.02	.68	2.67	0.90	.58
	Barmann	9.61	2.50	.56	7.50	2.07	.58	2.33	0.89	.44
	Begleitung	9.70	2.25	.70	7.59	1.90	.70	2.46	0.94	.55
	Umzug	11.20	2.25	.63	9.00	1.96	.62	2.81	0.79	.51
	Geburtstag	10.21	2.21	.74	8.52	1.94	.75	2.51	0.81	.51
	Fitness	10.42	2.27	.74	8.16	1.91	.74	2.61	0.87	.52
	Cronbach Alpha			.91			.91			.84
Verantwortungsbewusstsein	Schanze	11.56	2.18	.35	8.17	1.90	.33	2.75	0.88	.31
	Beratungsstelle	12.59	1.81	.52	9.14	1.53	.50	2.64	0.66	.36
	Silvester	11.58	2.49	.48	8.62	2.09	.47	2.97	0.97	.41
	Subventionen	10.55	2.23	.42	7.27	1.93	.39	2.18	0.73	.33
	Nachhilfestunden	11.71	2.13	.52	8.34	1.86	.50	2.47	0.75	.37
	Kind	10.90	2.00	.47	7.99	1.62	.47	2.64	0.85	.38
	Wohnung	11.67	2.12	.42	8.17	1.89	.37	2.63	0.80	.29
	Bergtour	10.61	2.30	.41	8.41	1.85	.45	2.80	0.88	.38
	Ampel	11.70	3.07	.53	8.53	2.38	.53	2.92	0.93	.43
	Autofahren	13.29	2.10	.51	9.78	1.90	.49	3.01	0.84	.39
	Cronbach Alpha			.79			.78			.70

Anhang 7.7 Itemkennwerte der Skala Durchsetzungsfähigkeit der likert-skalierten Version des Leadership-Fragebogens (Datensatz 2007; *N* = 1'017)

		M	*SD*	Trennschärfe r_{it}					
Anzahl Item-Stämme x Anzahl Antwortalternativen				10 x 4	8 x 4	6 x 4	10 x 3	8 x 3	6 x 3
Lohnerhöhung	LS_02BU	2.97	0.81	.20			.21		
	LS_02AU	2.07	0.78	.03					
	LS_02D	3.07	0.85	.20			.21		
	LS_02C	2.12	0.94	.22			.22		
Fahrer	LS_04CU	3.03	0.83	.38	.39	.40	.37	.38	.37
	LS_04DU	1.83	0.89	.22	.26	.26			
	LS_04A	3.13	0.87	.41	.43	.44	.40	.41	.39
	LS_04B	2.65	0.95	.42	.44	.43	.43	.43	.40
Unterbruch	LS_09CU	2.53	0.82	.34	.34	.36	.33	.33	.34
	LS_09D	3.22	0.76	.10	.08	.08			
	LS_09B	2.14	0.75	.27	.26	.24	.26	.24	.22
	LS_09A	2.34	0.85	.33	.31	.30	.36	.36	.36
Aufräumen	LS_13CU	3.51	0.69	.16	.13	.14	.16	.15	.20
	LS_13AU	1.94	0.91	.20	.20	.21			
	LS_13D	3.34	0.73	.38	.34	.33	.37	.36	.39
	LS_13B	2.14	0.96	.44	.40	.39	.43	.42	.43
Waschküche	LS_15BU	2.34	0.81	.38	.37	.38	.37	.37	.39
	LS_15C	3.48	0.69	.25	.22	.21			
	LS_15A	2.31	0.82	.36	.36	.36	.37	.36	.37
	LS_15D	1.67	0.77	.28	.26	.24	.30	.28	.28
Geschirr	LS_19BU	3.14	0.77	.28			.28	.26	.31
	LS_19C	2.54	0.96	.03					
	LS_19A	3.10	0.76	.27			.28	.27	.29
	LS_19D	2.15	0.98	.40			.42	.41	.41
Musik	LS_21AU	2.96	0.82	.34	.33		.32		
	LS_21C	3.37	0.73	.14	.12				
	LS_21B	2.92	0.85	.33	.32		.33		
	LS_21D	2.17	1.01	.16	.16		.17		
Probleme	LS_22AU	2.12	0.73	.27	.29		.25	.26	
	LS_22D	2.52	0.88	.06	.06				
	LS_22B	2.56	0.80	.31	.32		.31	.32	
	LS_22C	1.81	0.74	.30	.31		.30	.31	
Auswärts	LS_24CU	2.84	0.82	.30	.31	.32	.30	.31	
	LS_24BU	2.22	0.83	.16	.18	.18			
	LS_24A	2.80	0.78	.23	.24	.21	.23	.23	
	LS_24D	2.27	0.89	.37	.39	.36	.37	.39	
Arbeit	LS_27AU	2.92	0.84	.31	.32	.35	.27	.28	.31
	LS_27DU	2.57	0.97	.38	.41	.42			
	LS_27B	2.86	0.82	.38	.41	.40	.34	.35	.34
	LS_27C	2.44	0.94	.40	.40	.38	.38	.37	.35
Cronbach Alpha				.81	.80	.78	.81	.79	.77

Anhang 7.8 Itemkennwerte der Skala Kontaktfähigkeit der likert-skalierten Version des Leadership-Fragebogens (Datensatz 2007; *N* = 1'017)

		M	*SD*	Trennschärfe r_{it}					
Anzahl Item-Stämme x Anzahl Antwortalternativen				10 x 4	8 x 4	6 x 4	10 x 3	8 x 3	6 x 3
Nachbarn	LS_03BU	3.25	0.77	.49	.49	.49	.50	.50	
	LS_03CU	2.42	0.75	.43	.43	.43			
	LS_03A	3.35	0.76	.58	.57	.56	.60	.60	
	LS_03D	2.23	1.00	.48	.47	.47	.48	.47	
Zugfahrt	LS_06DU	3.15	0.87	.58	.57	.57	.59	.58	.58
	LS_06BU	2.16	0.81	.42	.43	.44			
	LS_06C	2.88	0.88	.58	.57	.56	.60	.59	.60
	LS_06A	2.20	0.84	.62	.61	.61	.61	.60	.60
Kurs	LS_08DU	3.14	0.88	.53			.54		
	LS_08BU	2.48	0.79	.24					
	LS_08A	2.93	0.82	.57			.56		
	LS_08C	1.53	0.68	.29			.28		
Schultag	LS_10DU	2.94	0.88	.48			.48		
	LS_10A	2.90	0.72	.17					
	LS_10C	3.09	0.69	.35			.38		
	LS_10B	2.56	0.89	.55			.55		
Flugzeug	LS_14CU	3.05	0.76	.52	.52		.52	.52	.53
	LS_14B	3.38	0.61	.23	.21				
	LS_14A	2.67	0.81	.49	.49		.52	.52	.53
	LS_14D	2.51	0.97	.62	.60		.61	.60	.60
Barmann	LS_16DU	2.51	0.91	.46	.47		.45	.45	
	LS_16BU	2.11	0.84	.26	.26				
	LS_16C	2.71	0.83	.65	.66		.65	.65	
	LS_16A	2.28	0.84	.39	.39		.39	.39	
Begleitung	LS_18AU	2.26	0.74	.29	.29	.32	.27	.27	.28
	LS_18CU	2.10	0.74	.33	.34	.34			
	LS_18D	2.79	0.88	.70	.70	.70	.68	.68	.67
	LS_18B	2.54	0.84	.66	.66	.64	.66	.66	.65
Umzug	LS_25AU	3.19	0.81	.46	.46	.46	.48	.48	.47
	LS_25DU	2.20	0.85	.25	.26	.27			
	LS_25B	3.16	0.75	.49	.48	.48	.52	.52	.51
	LS_25C	2.65	0.80	.64	.64	.64	.65	.64	.64
Geburtstag	LS_29BU	3.17	0.77	.46	.46	.48	.47	.47	.47
	LS_29AU	1.69	0.61	.31	.32	.33			
	LS_29D	2.64	0.88	.64	.64	.63	.63	.63	.62
	LS_29C	2.71	0.80	.68	.68	.66	.69	.68	.68
Fitness	LS_30DU	2.85	0.84	.42	.42	.42	.44	.44	.43
	LS_30CU	2.26	0.74	.38	.40	.39			
	LS_30B	2.59	0.77	.69	.70	.69	.69	.70	.70
	LS_30A	2.72	0.79	.68	.68	.65	.69	.69	.68
Cronbach Alpha				.93	.92	.91	.93	.92	.91

Anhang 7.9 Itemkennwerte der Skala Verantwortungsbewusstsein der likert-skalierten Version des Leadership-Fragebogens (Datensatz 2007; *N* = 1'017)

		M	*SD*	Trennschärfe r_{it}					
Anzahl Item-Stämme x Anzahl Antwortalternativen				10 x 4	8 x 4	6 x 4	10 x 3	8 x 3	6 x 3
Schanze	LS_01CU	2.56	.93	.25			.26		
	LS_01A	3.39	.77	.20					
	LS_01D	3.24	.88	.23			.20		
	LS_01B	2.38	.96	.28			.30		
Beratungsstelle	LS_05AU	3.44	.71	.44	.43	.45	.42	.42	.43
	LS_05D	3.45	.71	.23	.24	.24			
	LS_05C	3.52	.64	.43	.42	.40	.41	.39	.38
	LS_05B	2.19	.87	.25	.26	.25	.26	.25	.26
Silvester	LS_07BU	3.08	.89	.46	.47	.49	.45	.46	.49
	LS_07C	2.96	.91	.25	.27	.29			
	LS_07A	2.80	.90	.42	.45	.46	.41	.41	.45
	LS_07D	2.74	.91	.34	.35	.34	.37	.38	.37
Subventionen	LS_11AU	2.68	.85	.27	.26	.27	.27	.29	.28
	LS_11B	3.29	.67	.31	.31	.33			
	LS_11C	2.84	.88	.45	.46	.45	.43	.43	.44
	LS_11D	1.74	.85	.26	.27	.27	.29	.29	.30
Nachhilfestunden	LS_12DU	2.49	.93	.28	.25	.24	.30	.28	.26
	LS_12B	3.38	.68	.26	.25	.25			
	LS_12C	3.23	.79	.47	.47	.47	.44	.44	.43
	LS_12A	2.61	.87	.43	.43	.42	.42	.41	.40
Kind	LS_17DU	2.35	.98	.10	.09		.12		
	LS_17B	2.90	.77	.28	.30				
	LS_17A	3.10	.76	.43	.44		.42		
	LS_17C	2.55	.88	.37	.38		.36		
Wohnung	LS_20BU	2.56	.84	.28	.26		.31	.29	
	LS_20C	3.50	.62	.32	.32				
	LS_20D	3.18	.78	.43	.42		.39	.40	
	LS_20A	2.43	.90	.25	.23		.27	.26	
Bergtour	LS_23DU	2.88	.96	.21			.22	.20	
	LS_23AU	2.20	.88	.16					
	LS_23C	3.03	.85	.40			.38	.37	
	LS_23B	2.50	.89	.39			.39	.39	
Ampel	LS_26CU	3.04	.99	.53	.52	.53	.51	.51	.50
	LS_26AU	3.16	.93	.50	.49	.50			
	LS_26B	3.22	.93	.46	.46	.46	.43	.43	.42
	LS_26D	2.27	1.07	.45	.47	.45	.47	.46	.46
Autofahren	LS_28CU	3.39	.73	.45	.43	.43	.43	.45	.42
	LS_28D	3.51	.64	.19	.20	.18			
	LS_28A	3.48	.72	.48	.50	.50	.46	.47	.47
	LS_28B	2.91	.94	.39	.40	.40	.41	.42	.42
Cronbach Alpha				.86	.85	.83	.84	.83	.81

Anhang 7.10 Verteilungskennwerte der zwei Versionen des Leadership-Fragebogens in den zwei Stichproben „Stellungspflichtige" und „Studenten"

			min	max	*M*	*SD*	Schiefe	Exzess
Forced-Choice	Stellungspflichtige	Durchsetzungsfähigkeit	10	40	24.13	3.69	.22	.58
		Kontaktfähigkeit	10	40	26.18	5.35	-.50	.00
		Verantwortungsbewusstsein	10	39	27.02	4.37	-.61	.51
		Durchsetzungsfähigkeit	*16*	*37*	*23.99*	*3.80*	*.55*	*1.14*
		Kontaktfähigkeit	*10*	*37*	*26.00*	*5.87*	*-.47*	*-.20*
		Verantwortungsbewusstsein	*11*	*37*	*27.50*	*3.99*	*-.71*	*2.03*
	Studenten	Durchsetzungsfähigkeit	15	34	23.31	3.78	.29	.01
		Kontaktfähigkeit	11	34	24.51	4.50	-.07	-.29
		Verantwortungsbewusstsein	16	34	26.51	3.24	-.26	.32
likert-skaliert	Stellungs-pflichtige	Durchsetzungsfähigkeit	66	147	104.16	11.44	.04	.30
		Kontaktfähigkeit	53	147	105.92	16.72	-.22	-.23
		Verantwortungsbewusstsein	68	152	116.16	13.28	-.50	.47
	Studenten	*Durchsetzungsfähigkeit*[1)]	*56.00*	*120.00*	*95.57*	*12.53*	*-.48*	*.49*
		Kontaktfähigkeit	*51.33*	*139.33*	*94.90*	*16.79*	*.00*	*.48*
		Verantwortungsbewusstsein	*84.67*	*134.67*	*109.40*	*10.15*	*-.12*	*-.28*
		Durchsetzungsfähigkeit[2)]	84	180	143.35	18.80	-.48	.49
		Kontaktfähigkeit	77	209	142.35	25.19	.00	.48
		Verantwortungsbewusstsein	127	202	164.10	15.22	-.12	-.28

Anmerkung. *N* = 19'801 (Stellungspflichtige, Forced-Choice), *N* = 100 (*kursiv gesetzt*: Zufallsstichprobe aus dem Forced-Choice-Datensatz der Stellungspflichtigen), *N* = 1'017 (Stellungspflichtige, likert-skaliert), *N* = 100 (Studenten). [1)] Um die Vergleichbarkeit herzustellen, habe ich die sechsstufige Antwortskala der Papier-und-Bleistift-Version auf die vierstufige der Computerversion transformiert. [2)] Kennwerte der sechsstufigen Antwortskala.

Anhang 7.11 Beispiel-Item der likert-skalierten Version des bei der Datenerhebung bei den Psychologie-Studierenden eingesetzten Papier-und-Bleistift-Tests

1. Zelten

Mit ein paar Kollegen verbringen Sie eine Woche Ferien auf einem Zeltplatz am Mittelmeer. Es kommen immer mehr Leute und langsam wird es eng.

	trifft überhaupt nicht zu	trifft meist nicht zu	trifft eher nicht zu	trifft eher zu	trifft meist zu	trifft vollständig zu
Sie beschliessen, sich einen anderen Zeltplatz zu suchen, wenn es nicht besser wird.	☐	☐	☐	☐	☐	☐
Sie haben gerne Leute um sich herum und geniessen die spontanen Bekanntschaften auf dem Zeltplatz.	☐	☐	☐	☐	☐	☐
Sie brauchen auch mal Ihre Ruhe. Deshalb ziehen Sie sich öfters an einen ruhigeren Ort zurück.	☐	☐	☐	☐	☐	☐
Sie geniessen den Rummel: Je mehr Leute um Sie herum, desto besser. Sonst wäre es ja langweilig.	☐	☐	☐	☐	☐	☐

Anhang 7.12 Überprüfung der Wertequadrate bei der likert-skalierten Version des Leadership-Fragebogens (Datensatz 2007; *N* = 1'017)

		Anteil der Probanden mit jeweils höchster Antwortausprägung im betreffenden Wertequadranten (WQ)							
		WQ 1	WQ 1-2	WQ 2	WQ 2-3	WQ 3	WQ 3-4	WQ 4	unein-deutig
Durchsetzungsfähigkeit	Lohnerhöhung	2.6	2.0	18.3	19.7	23.7	3.7	.6	29.5
	Fahrer	1.4	1.3	28.8	11.5	15.2	9.9	.3	31.6
	Unterbruch	5.2	1.7	23.0	4.5	2.3	.4	.1	62.8
	Aufräumen	.3	.5	23.1	26.5	26.8	4.2	.4	18.2
	Waschküche	8.5	7.8	46.3	9.5	2.4	.1	0	25.5
	Geschirr	2.2	.5	12.6	18.1	26.5	3.7	.8	35.7
	Musik	1.4	1.2	29.0	26.9	7.3	1.1	0	33.1
	Probleme	13.3	5.4	9.2	7.4	9.5	.2	.2	54.8
	Auswärts	3.0	1.8	18.5	11.7	16.4	2.4	.9	45.3
	Arbeit	1.7	2.8	13.5	8.1	18.4	3.1	.4	52.1
Kontaktfähigkeit	Nachbarn	1.9	2.0	9.4	12.4	44.9	7.3	1.4	20.7
	Zugfahrt	3.4	2.0	25.1	16.0	23.5	5.4	1.4	23.2
	Kurs	4.4	2.1	16.7	14.2	41.7	1.0	.1	19.9
	Schultag	1.9	1.0	9.9	15.8	14.9	5.5	.6	50.3
	Flugzeug	2.6	2.7	32.5	14.3	2.8	.9	.5	43.9
	Barmann	6.7	4.6	18.4	8.9	9.9	3.5	.5	47.4
	Begleitung	4.2	2.3	14.2	4.5	10.0	3.2	.3	61.3
	Umzug	2.6	1.3	15.3	14.5	24.9	7.5	.8	33.2
	Geburtstag	.4	.4	28.4	7.5	5.7	6.8	.5	50.3
	Fitness	3.1	.7	17.2	3.1	4.0	7.4	1.7	62.7
Verantwortungsbewusstsein	Schanze	4.8	2.6	14.0	28.1	12.3	1.3	.2	36.8
	Beratungsstelle	.7	.3	19.3	41.3	20.1	1.8	.2	16.4
	Silvester	1.9	.8	14.9	16.1	9.0	2.2	2.0	53.1
	Subventionen	6.8	4.9	28.1	27.5	7.8	.7	.1	24.1
	Nachhilfestunden	4.2	2.2	17.5	18.6	11.4	2.8	0	43.4
	Kind	2.9	1.3	7.9	10.3	15.5	1.4	.1	60.6
	Wohnung	5.4	2.9	21.9	25.0	7.4	1.4	.3	35.7
	Bergtour	1.8	1.6	18.5	13.6	17.0	4.1	1.3	42.2
	Ampel	5.2	4.8	1.2	1.5	38.5	2.7	.1	46.0
	Autofahren	.9	.1	18.4	25.0	10.8	8.5	.5	35.9

Anhang 7.13 Bekanntheitsgrad der Aussagen im Leadership-Fragebogen (Datensatz 2009; *N* = 98-100)

Dimension	Item	Situation erlebt		sich hineinversetzen	
		M	*SD*	*M*	*SD*
Durchsetzungsfähigkeit	Disco	.65	.48	.94	.24
	Lohnerhöhung	.20	.40	.83	.38
	Fahrer	.50	.50	.96	.20
	Unterbruch	.93	.26	.99	.10
	Zugreise	.65	.48	.93	.26
	Aufräumen	.47	.50	.92	.27
	Waschküche	.35	.48	.87	.34
	Geschirr	.34	.48	.89	.31
	Musik	.44	.50	.95	.22
	Probleme	.63	.49	.93	.26
	Auswärts	.63	.49	.94	.24
	Arbeit	.53	.50	.95	.22
	Schülerzeitung	.07	.26	.77	.42
Mittelwert Durchsetzungsfähigkeit		.49	.21	.91	.14
Kontaktfähigkeit	Zelten	.55	.50	.97	.17
	Nachbarn	.52	.50	.98	.14
	Zugfahrt	.91	.29	.97	.17
	Kurs	.84	.37	1.00	.00
	Schultag	.93	.26	.96	.20
	Flugzeug	.74	.44	.98	.14
	Barmann	.46	.50	.89	.31
	Begleitung	.88	.33	.99	.10
	Party	.56	.50	.91	.29
	Allein	.57	.50	.96	.20
	Umzug	.38	.49	.92	.27
	Geburtstag	.80	.40	.98	.14
	Fitness	.62	.49	.97	.17
Mittelwert Kontaktfähigkeit		.67	.19	.96	.07
Verantwortungsbewusstsein	Schanze	.34	.48	.91	.29
	Unstimmigkeiten	.49	.50	.92	.27
	Beratungsstelle	.42	.50	.91	.29
	Silvester	.29	.46	.88	.33
	Subventionen	.22	.42	.82	.39
	Nachhilfestunden	.37	.49	.93	.26
	Kind	.31	.46	.86	.35
	Malediven	.06	.24	.77	.42
	Wohnung	.30	.46	.92	.27
	Bergtour	.11	.31	.80	.40
	Ampel	.74	.44	.99	.10
	Autofahren	.55	.50	.96	.20
	Mädchen	.35	.48	.93	.26
Mittelwert Verantwortungsbewusstsein		.35	.19	.89	.15

Anhang 7.14 Akzeptanz-Fragebogen 1

		Item-Format					
		likert-skaliert		Leader		Forced-Choice	
		M	*SD*	*M*	*SD*	*M*	*SD*
Layout	Cronbach Alpha = .82	4.03	1.06	4.79	1.25	3.11	1.41
1	Das Fragebogenformat empfinde ich ansprechend.	4.17	1.24	4.89	1.33	3.30	1.54
11	Die Art der Aufmachung der Aufgaben hat mir gut gefallen.	3.89	1.22	4.69	1.43	2.93	1.55
Verständlichkeit	Cronbach Alpha = .66	4.83	1.15	4.96	1.10	3.94	1.35
2	Die Aussagen sind für mich verständlich formuliert.	4.97	1.24	5.18	1.07	4.15	1.37
7	Ich musste einige Aussagen zweimal lesen, um sie zu verstehen. (-)	4.69	1.45	4.74	1.47	3.73	1.83
Erleben	Cronbach Alpha = .73	4.07	1.08	4.49	1.18	3.49	1.18
6	Das Ausfüllen des Fragebogens hat mir Spass gemacht.	3.62	1.42	4.18	1.55	2.99	1.50
14	Ich habe es als unangenehm empfunden, auf die Aussagen zu antworten. (-)	4.84	1.36	5.02	1.27	4.33	1.52
15	Ich fand es interessant, den Fragebogen zu beantworten.	3.75	1.46	4.28	1.46	3.16	1.54
Augenscheinvalidität	Cronbach Alpha = .60	4.24	.92	3.97	.90	3.98	1.01
4	Die Aussagen widerspiegeln Anforderungen, die auch im Berufsleben von einer Führungsperson gefordert werden.	4.41	1.09	3.75	1.17	4.12	1.19
9	Ich denke, dass anhand meiner Antworten Aussagen über mein Verhalten gemacht werden können.	4.10	1.51	4.04	1.45	3.93	1.36
13	Ich kann mir gut vorstellen, welche Persönlichkeitseigenschaften mit dem Fragebogen untersucht werden.	4.21	1.15	4.12	1.11	3.89	1.25
Alltäglichkeit	Cronbach Alpha = .68	4.47	.96	4.57	.98	3.42	1.03
3	Ich kenne den Inhalt der Aussagen aus meinem Alltagsleben.	4.41	1.18	4.45	1.27	3.97	1.28
8	Ich konnte mich schnell für eine Antwort entscheiden.	4.68	1.28	4.53	1.33	2.66	1.49
12	Ich konnte mich gut in die beschrieben Situationen versetzen.	4.32	1.22	4.73	1.31	3.63	1.33
Privatsphäre	Cronbach Alpha = .68	4.52	1.14	4.71	1.19	4.56	1.17
5	Die Aufgaben dringen zu tief in meine Privatsphäre ein. (-)	4.70	1.33	4.89	1.28	4.61	1.33
10	Ich fühle mich durch das Beantworten des Fragebogens durchleuchtet. (-)	4.34	1.34	4.53	1.39	4.50	1.38

Anmerkung. $N_{Cronbach\ Alpha}$ = 393; $N_{Mittelwerte}$ = 131. Die Antwortskala reicht von 1 bis 6.

Anhang 7.15a Einfaktorielle Varianzanalysen und Post-hoc-Vergleiche der Akzeptanzunterschiede der drei Itemformate

		Post-hoc-Vergleiche nach Tukey resp. Games-Howell		
Skala	einfaktorielle Varianzanalysen	Leader – Likert	Leader – F-C	Likert – F-C
Layout[1)]	$F(2, 256.29) = 51.19$, $p < .001$, $\omega = .48$	$p < .001$	$p < .001$	$p < .001$
Verständlichkeit[1)]	$F(2, 258.19) = 24.80$, $p < .001$, $\omega = .35$	$p = .60$	$p < .001$	$p < .001$
Erleben	$F(2, 390) = 24.77$, $p < .001$, $\omega = .33$	$p < .01$	$p < .001$	$p < .001$
Augenscheinvalidität	$F(2, 390) = 3.48$, $p < .05$, $\omega = .11$	$p = .05$	$p = .99$	$p = .07$
Alltäglichkeit	$F(2, 390) = 54.17$, $p < .001$, $\omega = .46$	$p = .70$	$p < .001$	$p < .001$
Privatsphäre	$F(2, 390) = .95$, $p = .39$, $\omega = .02$	$p = .40$	$p = .54$	$p = .97$

Anmerkung. $N = 131$. F-C = Forced-Choice.
[1)] Verfahren nach Welch resp. Games-Howell auf Grund inhomogener Varianzen.

Anhang 7.15b Rangvarianzanalyse und Mittelwertsvergleiche der Rangierung der drei Itemformate

		Mittelwertsvergleiche zwischen den einzelnen Itemformaten (Vorzeichenrangtest von Wilcoxon)		
	Rangvarianzanalyse (Friedmann)	Leader – Likert	Leader – Forced-Choice	Likert – Forced-Choice
1	$\chi^2(2) = 93.30$, $p < .001$	$T = 2'864$, $p < .017$, $r = -.16$	$T = 682.5$, $p < .001$, $r = -.52$	$T = 1'269$, $p < .001$, $r = -.42$
2	$\chi^2(2) = 152.85$, $p < .001$	$T = 882$, $p < .001$, $r = -.50$	$T = 147$, $p < .001$, $r = -.61$	$T = 1'587$, $p < .001$, $r = -.39$
3	$\chi^2(2) = 115.15$, $p < .001$	$T = 2'335.5$, $p < .001$, $r = -.25$	$T = 1'075.5$, $p < .001$, $r = -.46$	$T = 439.5$, $p < .001$, $r = -.56$
4	$\chi^2(2) = 106.34$, $p < .001$	$T = 3'713$, $p = .92$, $r = -.00$	$T = 772$, $p < .001$, $r = -.50$	$T = 609.5$, $p < .001$, $r = -.53$
5	$\chi^2(2) = 52.87$, $p < .001$	$T = 3'329$, $p = .24$, $r = -.08$	$T = 1'599$, $p < .001$, $r = -.36$	$T = 1'681.5$, $p < .001$, $r = -.35$
6	$\chi^2(2) = 2.80$, $p = .25$	$T = 3'106.5$, $p = .08$, $r = -.11$	$T = 3'729$, $p = .95$, $r = -.00$	$T = 3'309$, $p = .24$, $r = -.08$

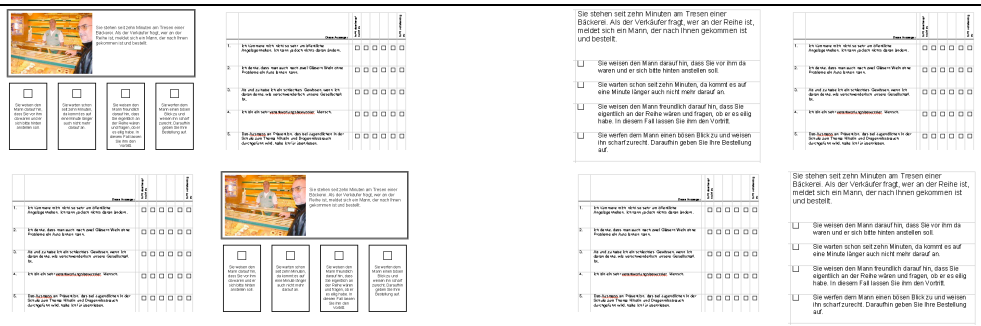
Anmerkung. $N = 122$. 1-6: Fragen des Rating-Fragebogens (Wortlaut siehe Tabelle oben). Das Signifikanzniveau bei den Einzelvergleichen ist Bonferroni-korrigiert ($p < .017$).

Anhang 7.16 Akzeptanz-Fragebogen 2

		Item-Format					
		Leader Bild		Leader Text		Likert	
		M	*SD*	*M*	*SD*	*M*	*SD*
Layout	Cronbach Alpha = .76	4.33	1.14	3.88	1.33	3.27	1.27
1	Dieses Fragebogenformat empfinde ich ansprechend.	4.31	1.37	3.90	1.53	3.34	1.49
10	Die Art der Aufmachung der Aufgaben hat mir gut gefallen.	4.36	1.27	3.86	1.41	3.19	1.36
Verständlichkeit	Cronbach Alpha = .58	5.02	.85	4.98	.92	4.83	.99
2	Die Aussagen in diesem Fragebogen sind für mich verständlich formuliert.	5.24	.98	5.17	1.06	5.00	1.21
7	Ich musste einige Aussagen zweimal lesen, um sie zu verstehen. (-)	4.96	1.27	4.95	1.35	4.67	1.56
13	Ich habe beim Ausfüllen des Fragebogens nicht recht gewusst, was von mir verlangt wird. (-)	4.88	1.25	4.81	1.25	4.82	1.23
Erleben	Cronbach Alpha = .88	3.67	1.42	3.45	1.50	2.95	1.40
8	Das Ausfüllen des vorhergehenden Fragebogens hat mir Spass gemacht.	3.58	1.50	3.36	1.56	2.83	1.46
14	Ich fand es interessant, den vorhergehenden Fragebogen zu beantworten.	3.76	1.52	3.53	1.59	3.06	1.53
Augenscheinvalidität	Cronbach Alpha = .56	4.07	1.03	3.95	1.04	4.18	1.05
4	Dieser Fragebogen ist leicht durchschaubar.	4.04	1.35	4.01	1.31	4.12	1.35
6	Es wäre sehr einfach, sich in diesem Fragebogen besser darzustellen als man wirklich ist.	3.97	1.51	3.87	1.54	4.27	1.51
12	Ich kann mir gut vorstellen, welche Persönlichkeitseigenschaften mit diesem Fragebogen untersucht werden.	4.19	1.31	3.96	1.45	4.15	1.45
Alltäglichkeit	Cronbach Alpha = .63	4.30	1.08	4.19	1.21	4.09	1.11
3	Ich kenne den Inhalt der Aussagen im Fragebogen aus meinem Alltagsleben.	4.08	1.31	3.99	1.41	4.16	1.35
11	Ich konnte mich gut in die im Fragebogen beschriebenen Situationen versetzen.	4.51	1.22	4.38	1.36	4.01	1.30
Privatsphäre	Cronbach Alpha = .56	4.82	1.00	4.84	.99	4.68	1.15
5	Die Aussagen im Fragebogen dringen zu tief in meine Privatsphäre ein. (-)	5.09	1.12	5.04	1.21	4.83	1.30
9	Ich fühle mich durch das Beantworten des Fragebogens durchleuchtet. (-)	4.56	1.31	4.65	1.28	4.53	1.40

Anmerkung. $N_{Cronbach\ Alpha}$ = 1'370; $N_{Leader\ Bild}$ =337; $N_{Leader\ Text}$ = 343; N_{Likert} = 690. Die Antwortskala reicht von 1 bis 6.

Anhang 7.17 Präferenzurteile im Akzeptanz-Vergleichs-Fragebogen 2



	n = 166	*n* = 174	∅	*n* = 171	*n* = 176	∅
1	**59.04**	9.20	***67.65***	**53.22**	17.05	***62.25***
	18.07	14.94	*16.47*	15.20	11.93	*13.54*
	22.89	**75.86**	*15.88*	31.58	**71.02**	*24.21*
2	**77.71**	8.05	***82.06***	**49.71**	17.05	***57.06***
	7.23	5.75	*6.47*	19.30	18.75	*19.02*
	15.06	**86.21**	*11.47*	30.99	**64.20**	*23.92*
3	**30.72**	36.21	***39.41***	**31.58**	38.07	***40.63***
	15.66	16.09	*15.88*	18.13	12.50	*15.27*
	53.61	**47.70**	*44.71*	50.29	**49.43**	*44.09*
4	**39.76**	18.97	***48.82***	**42.69**	26.14	***46.11***
	25.30	23.56	*24.41*	26.32	24.43	*25.36*
	34.94	**57.47**	*26.76*	30.99	**49.43**	*28.53*
5	**46.39**	16.67	***54.41***	**47.37**	18.18	***56.48***
	27.11	21.26	*24.12*	23.39	16.48	*19.88*
	26.51	**62.07**	*21.47*	29.24	**65.34**	*23.63*
6	**32.53**	23.56	***37.94***	**32.16**	22.73	***41.21***
	34.34	33.33	*33.82*	26.32	27.27	*26.80*
	33.13	**43.10**	*28.24*	41.52	**50.00**	*31.99*
∅	**47.69**	18.77	***55.05***	**42.79**	23.20	***50.62***
	21.29	19.16	*20.20*	21.44	18.56	*19.98*
	31.02	**62.07**	*24.75*	35.77	**58.24**	*29.39*

Nr.	Item-Text
1	Wenn Sie nochmals an einer Befragung teilnehmen müssten, welches Frageformat würden Sie bevorzugen?
2	Welches Fragebogenformat hat Ihnen vom Layout her am besten gefallen?
3	Bei welchem Frageformat konnten Sie sich am schnellsten für eine Antwort entscheiden?
4	Welcher Fragebogentyp enthielt die am leichtesten zu verstehenden Aussagen?
5	Welcher Fragebogentyp enthält Aussagen, die Sie am ehesten ansprechen, da Sie sie mit persönlichen Erfahrungen verbinden können?
6	Mit welchem Frageformat können Ihrer Meinung nach am ehesten Aussagen über die Persönlichkeit eines Menschen gesammelt werden?

Anhang 7.18 Vergleich der zwei Versionen des Akzept!-Fragebogens

Die Akzept!-Skalen sind nicht zur Veröffentlichung freigegeben. Die Skalen können für den Einsatz in Studien bei M. Kersting angefordert werden.

Anhang 7.18 Vergleich der zwei Versionen des Akzept!-Fragebogens (Fortsetzung)

Die Akzept!-Skalen sind nicht zur Veröffentlichung freigegeben. Die Skalen können für den Einsatz in Studien bei M. Kersting angefordert werden.

Anhang 7.18 Vergleich der zwei Versionen des Akzept!-Fragebogens (Fortsetzung)

Die Akzept!-Skalen sind nicht zur Veröffentlichung freigegeben. Die Skalen können für den Einsatz in Studien bei M. Kersting angefordert werden.

Anhang 7.19 Ablauf des kognitiven Pretests der Akzept!-Skalen

Das Interview wird an einem ruhigen, störungsfreien Ort durchgeführt. Der Interviewer begrüsst den Stellungspflichtigen und erklärt diesem das Ziel und den Ablauf des Interviews:

„Sie haben heute am Computer mehrere Testverfahren bearbeitet. Ich lege Ihnen jetzt einen Ausschnitt aus einem Fragebogen vor, mit welchem die Akzeptanz dieser Tests bei den Stellungspflichtigen erhoben werden kann. Es geht jetzt aber nicht einfach darum, dass Sie diese Fragen beantworten, sondern dass Sie eine aktive Rolle bei der Entwicklung dieses Fragebogens einnehmen, indem Sie mir jeweils berichten, welche Gedanken Ihnen bei der Beantwortung der einzelnen Fragen dieses Akzeptanz-Fragebogens durch den Kopf gegangen sind. Sie sind für mich in diesem Sinne ein externer Experte für die Entwicklung des Fragebogens. Ich werde Ihre Aussagen auf Tonband aufzeichnen, damit ich diese später auswerten und analysieren kann. Haben Sie noch Fragen zu diesem Interview?"

„Die Fragen des Akzeptanz-Fragebogens beziehen sich auf den Intelligenz-Test / Persönlichkeitstest, den Sie am Computer bearbeitet haben. Als Erinnerungshilfe habe ich Ihnen ein paar Beispiele der Aufgaben / Aussagen, welche in diesem Test vorgekommen sind. Können Sie sich daran erinnern, diesen Test am Computer bearbeitet zu haben?"

„Lesen Sie nun die erste Frage für sich durch und beantworten Sie diese, indem Sie auf der sechsstufigen Skala an der Ihrer Ansicht entsprechenden Ausprägung ein Kreuz setzen."

„Warum haben Sie das Kreuz bei der Ausprägung XY gesetzt?"

„Welche Gedanken sind Ihnen durch den Kopf gegangen, als Sie sich überlegt haben, bei welcher Ausprägung Sie das Kreuz setzen sollen?"

„Wie schwer fiel Ihnen die Beantwortung dieser Frage? Geben Sie mir dies auf einer Skala von 1 = einfach bis 6 = schwierig an." *Nachfrage bei Antworten zwischen 4 und 6:* „Was genau machte Ihnen bei der Beantwortung der Frage Mühe?"

„Gibt es etwas, was Sie in dieser Frage nicht verstanden haben oder Ihnen auf eine andere Weise bei der Beantwortung Mühe bereitet hat?"

Nachdem der Stellungspflichtige alle Fragen bearbeitet hat, stellt der Interviewer noch die Abschlussfrage:

„Zum Abschluss möchte ich Ihnen noch die Gelegenheit geben, sich gesamthaft zu diesem Ausschnitt des Akzeptanz-Fragebogens zu äussern. Wie haben Sie die Beantwortung der Akzeptanz-Fragen erlebt? Was ist Ihnen besonders aufgefallen? Was fanden Sie persönlich schwierig?"

Anhang 7.20 Itemkennwerte des Akzept!-Fragebogens

		Akzept!-P Leadership-Fragebogen		Akzept!-P Persönlichkeitstest		Akzept!-L Intelligenztest 95	
		M	*SD*	*M*	*SD*	*M*	*SD*
Kontrollierbarkeit		5.36	.76	5.31	.87	5.01	.97
1	7	5.53	1.12	5.45	1.17	5.17	1.23
9	15	5.58	.98	5.70	.83	5.30	1.08
13	9	5.00	1.23	5.05	1.23	4.87	1.49
19	1	5.32	1.12	5.25	1.07	4.69	1.38
Messqualität		3.65	.97	3.81	1.10	3.30	.98
2	2	3.86	1.26	3.88	1.32	3.56	1.33
8	6	3.46	1.24	3.74	1.28	3.22	1.26
12	11	3.58	1.36	3.69	1.37	3.04	1.31
18	16	3.68	1.31	3.94	1.33	3.40	1.24
Augenscheinvalidität		3.85	.94	3.97	.94	3.13	1.02
3	3	4.22	1.28	4.44	1.23	3.54	1.38
7	5	3.61	1.47	3.94	1.42	3.36	1.47
17	10	3.35	1.44	3.45	1.46	2.51	1.33
20	13	4.20	1.35	4.06	1.30	3.08	1.41
Privatsphäre		4.98	1.09	4.81	.96		
5	–	4.55	1.57	4.42	1.57		
11	–	5.28	1.23	5.05	1.18		
16	–	5.32	1.22	5.34	1.08		
21	–	4.77	1.48	4.44	1.62		
unverfälschte Antw.		4.97	1.09	5.02	.90		
4	–	5.02	1.58	4.95	1.60		
14	–	4.67	1.71	4.86	1.51		
22	–	4.89	1.54	4.97	1.37		
23	–	5.31	1.17	5.29	1.12		
Antwortfreiheit		4.15	1.05	4.25	1.07		
6	–	4.07	1.36	4.16	1.30		
10	–	3.92	1.47	3.94	1.63		
15	–	4.45	1.40	4.64	1.29		
Belastungsfreiheit						4.46	.91
–	4					4.64	1.32
–	8					4.47	1.40
–	12					4.85	1.10
–	14					3.88	1.50
Zusatzfragen							
24	17	2.88	1.51	2.86	1.45	3.43	1.61
	20					2.88	1.32
25	18	2.93	1.48	3.01	1.50	3.17	1.50
	21					2.85	1.33
26	19	3.41	1.44	3.27	1.49	3.58	1.53
	22					2.98	1.44
27		4.35	1.33	3.92	1.43		
Gesamtbeurteilung		4.28	.88	4.46	.73	4.15	.91
Selbstbeurteilung						4.30	.78

Anmerkung. N_{Leader} = 207-220, N_{Pers} = 163-178 resp. 73, N_{IQ} = *202-208 resp. 96*.

Anhang 7.21 Skalenkennwerte, Skalenreliabilitäten und Skalen-Interkorrelationen der drei Akzept!-Versionen

Leadership-Fragebogen	*M*	*SD*	1	2	3	4	5	6
1. Kontrollierbarkeit	5.36	.76	*.62*					
2. Messqualität	3.65	.97	.17	*.74*				
3. Augenscheinvalidität	3.85	.94	.24*	.64*	*.49*			
4. Wahrung der Privatsphäre	4.98	1.09	.43*	.27*	.38*	*.80*		
5. Intention zur unverfälschten Antwort	4.97	1.09	.32*	.22*	.22*	.39*	*.69*	
6. Antwortfreiheit	4.15	1.05	.32*	.51*	.42*	.27*	.23*	*.60*
7. Gesamtbeurteilung	4.28	.88	.16	.48*	.49*	.40*	.25*	.37*

Anmerkung. *N* = 207 - 220. *Bonferroni-korrigiertes Signifikanzniveau $p < .002$.
In kursiv sind die Skalenreliabilitäten nach Cronbach Alpha aufgeführt.

Persönlichkeits-Fragebogen	*M*	*SD*	1	2	3	4	5	6
1. Kontrollierbarkeit	5.31	.87	*.68*[1]					
2. Messqualität	3.81	1.10	.25*	*.85*				
3. Augenscheinvalidität	3.97	.94	.27*	.66*	*.52*			
4. Wahrung der Privatsphäre	4.81	.96	.27*	.19	.26*	*.64*		
5. Intention zur unverfälschten Antwort	5.02	.90	.48*	.10	.26*	.36*	*.51*	
6. Antwortfreiheit	4.25	1.07	.15	.56*	.47*	.29*	.23*	*.63*
7. Gesamtbeurteilung	4.46	.73	.35*	.34*	.39*	.21	.05	.15

Anmerkung. *N* = 163 – 178. *Bonferroni-korrigiertes Signifikanzniveau $p < .002$. [1)]*N* = 126.
In kursiv sind die Skalenreliabilitäten nach Cronbach Alpha aufgeführt.

Intelligenztest	*M*	*SD*	1	2	3	4
1. Kontrollierbarkeit	5.01	.97	*.74*			
2. Messqualität	3.30	.98	.12	*.76*		
3. Augenscheinvalidität	3.13	1.02	.16	.61*	*.64*	
4. Belastungsfreiheit	4.46	.91	.50*	.07	.24*	*.61*
5. Gesamtbeurteilung	4.15	.91	.36*	.44*	.42*	.27*

Anmerkung. *N* = 202 – 208. *Bonferroni-korrigiertes Signifikanzniveau $p < .005$.
In kursiv sind die Skalenreliabilitäten nach Cronbach Alpha aufgeführt.

Anhang 7.22 Itemkennwerte und Interkorrelationen der Zusatzfragen der drei Akzept!-Versionen

	Leadership-Fragebogen				Persönlichkeits-Fragebogen			
	M	*SD*	1	2	*M*	*SD*	1	2
1. Bearbeitung	2.90	1.41			2.94	1.40		
2. Darstellung	3.41	1.44	.59*		3.27	1.49	.70*	
3. Hineinversetzen	4.35	1.33	.34*	.47*	3.92	1.43	.36*	.51*

Anmerkung. N = 220 resp. 73. *Bonferroni-korrigiertes Signifikanzniveau $p < .017$.

Intelligenztest	*M*	*SD*	1	2	3
1. Bearbeitung hat Spass gemacht / unterhalt. (Verbal)	3.30	1.49			
2. Darstellung gefällt mir (Verbal)	3.58	1.53	.56*		
3. Bearbeitung hat Spass gemacht / unterhalt. (Figural)	2.86	1.24	.26*	.16*	
4. Darstellung gefällt mir (Figural)	2.98	1.44	.28*	.26*	.74*

Anmerkung. N = 96. *Bonferroni-korrigiertes Signifikanzniveau $p < .008$.

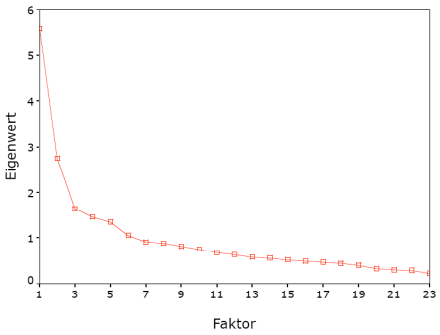
Anhang 7.23 Einfaktorielle Varianzanalysen und Post-hoc-Vergleiche respektive t-Tests der Akzeptanzunterschiede bei den drei Testverfahren

		Post-hoc-Vergleiche nach Tukey resp. Games-Howell		
Skala	einfaktorielle Varianzanalysen	Leader – Pers	Leader – IQ	Pers – IQ
Kontrollierbarkeit[1)]	$F(2, 387.15) = 9.28$, $p < .001$, $\omega = .17$	$p = .84$	$p < .001$	$p < .01$
Messqualität	$F(2, 601) = 12.71$, $p < .001$, $\omega = .19$	$p = .25$	$p < .01$	$p < .001$
Augenscheinvalidität	$F(2, 603) = 44.48$, $p < .001$, $\omega = .35$	$p = .41$	$p < .001$	$p < .001$
Privatsphäre[2)]	$t(396) = 1.57$, $p = .12$, $r = .08$			
unverfälschte Antwort[1,2)]	$t(395.81) = -.50$, $p = .62$, $r = .03$			
Antwortfreiheit[2)]	$t(396) = -.95$, $p = .34$, $r = .05$			
Gesamtbeurteilung	$F(2, 573) = 5.94$, $p < .01$, $\omega = .13$	$p = .10$	$p = .30$	$p < .01$
unterhaltsame Bearbeitung[1,3,4)]	$F(3, 202.58) = 3.14$, $p < .05$, $\omega = .12$	$p = 1.00$	$p < .05$ $p = 1.00$	$p = .08$ $p = 1.00$
schöne Darstellung[3,4)]	$F(3, 481) = 3.05$, $p < .05$, $\omega = .11$	$p = .90$	$p = .78$ $p = .08$	$p = .53$ $p = .57$
Hineinversetzen[3)]	$t(290) = 2.38$, $p < .05$, $r = .14$			

Anmerkung. N_{Leader} = 220, N_{Pers} = 178 resp. 73, N_{IQ} = 208 resp. 96.
Leader = Leadership-Fragebogen, Pers = Persönlichkeits-Fragebogen, IQ = Intelligenztest, IQv = Intelligenztest Verbalteil, IQf = Intelligenztest Figuralteil
[1)] Verfahren nach Welch resp. Games-Howell auf Grund inhomogener Varianzen.
[2)] Dimension beim Intelligenz-Test nicht erfasst.
[3)] Bei Pers und IQ Datensätze mit reduziertem N.
[4)] Bei den Post-hoc-Vergleichen bezieht sich die erste Zeile auf den IQv, die zweite auf den IQf. Post-hoc-Vergleich zwischen IQv und IQf: $p < .05$

Anhang 7.24a Faktorenanalyse der Akzept!-Skala zur Einstufung des Leadership-Fragebogens

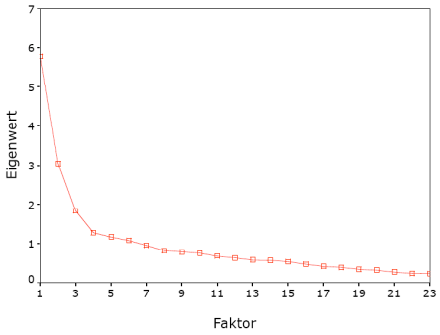
n	220
KMO	.82
Kleinster KMO-Wert (Grenzwert .50)	.67
Bartlett–Test	χ^2 = 1'556, *df* = 253, *p* < .001
Determinante der Korrelationsmatrix (Grenzwert .00001)	$\|R\|$ = 0.00039
Test nach Haitovsky	χ^2_H = 0.08, *df* = 253, *p* > .05
Anzahl Iterationen	6
Erklärte Varianz	55.57



	Faktor 1	Faktor 2	Faktor 3	Faktor 4	Faktor 5
Messqualität_3	**.77**	.01	.05	.17	.03
Messqualität_2	**.72**	.14	.06	.04	-.02
Messqualität_4	**.68**	.05	.02	-.03	.15
Messqualität_1	**.66**	-.03	.15	**.30**	-.21
Augenscheinvalidität_4	**.57**	.21	.11	-.10	.27
Antwortfreiheit_1	**.57**	-.03	.14	**.39**	.29
Augenscheinvalidität_2	**.54**	.21	-.04	.00	.21
Augenscheinvalidität_3	**.54**	**.36**	.16	**-.39**	.05
Augenscheinvalidität_1	**.49**	.09	-.10	**.45**	-.15
Privatsphäre_4	.16	**.75**	.15	.09	.04
Privatsphäre_2	.02	**.73**	.24	.29	.05
Privatsphäre_3	.05	**.72**	.16	**.40**	.04
Privatsphäre_1	.21	**.64**	.09	-.02	.03
unverfälschte Antwort_4	.12	.17	**.80**	.05	.07
unverfälschte Antwort_3	.02	.18	**.79**	.06	-.09
unverfälschte Antwort_1	-.08	.12	**.68**	.09	**.32**
unverfälschte Antwort_2	.26	.10	**.49**	.12	-.24
Kontrollierbarkeit_3	.15	.09	.00	**.68**	.08
Kontrollierbarkeit_4	.02	.26	.17	**.63**	.11
Kontrollierbarkeit_2	-.01	.24	**.34**	**.61**	.20
Antwortfreiheit_2	**.39**	-.08	-.07	-.04	**.61**
Kontrollierbarkeit_1	-.02	**.32**	-.07	.22	**.60**
Antwortfreiheit_3	.17	-.01	**.35**	.24	**.60**
Eigenwert	3.85	2.59	2.47	2.25	1.63
erklärte Varianz (%)	16.72	11.25	10.76	9.78	7.06

Anhang 7.24b Faktorenanalyse der Akzept!-Skala zur Einstufung des Persönlichkeits-Fragebogens

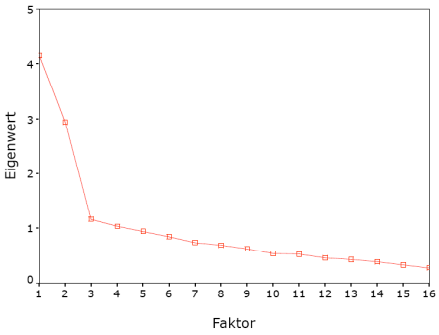
n	172
KMO	.82
Kleinster KMO-Wert (Grenzwert .50)	.69
Bartlett–Test	χ^2 = 1'409, *df* = 253, *p* < .001
Determinante der Korrelationsmatrix (Grenzwert .00001)	|*R*| = 0.00017
Test nach Haitovsky	χ^2_H = 0.03, *df* = 253, *p* > .05
Anzahl Iterationen	5
Erklärte Varianz	46.30



	Faktor 1	Faktor 2	Faktor 3
Messqualität_2	**.83**	.03	-.01
Messqualität_3	**.77**	.00	-.08
Messqualität_1	**.74**	.28	.09
Messqualität_4	**.70**	.21	-.06
Antwortfreiheit_1	**.69**	.05	.22
Augenscheinvalidität_2	**.64**	.16	.03
Augenscheinvalidität_3	**.60**	-.12	.08
Antwortfreiheit_2	**.57**	-.04	.17
Augenscheinvalidität_4	**.52**	**.33**	.06
Augenscheinvalidität_1	**.51**	**.33**	.08
Kontrollierbarkeit_2	.10	**.78**	.16
Kontrollierbarkeit_4	.05	**.74**	.21
Kontrollierbarkeit_3	.18	**.73**	.11
Kontrollierbarkeit_1	-.02	**.70**	-.08
unverfälschte Antwort_1	.08	**.44**	.07
unverfälschte Antwort_2	.12	**.38**	.19
Privatsphäre_2	-.04	.17	**.76**
Privatsphäre_3	-.03	.15	**.72**
Privatsphäre_4	.13	-.03	**.68**
unverfälschte Antwort_4	.03	**.35**	**.57**
Antwortfreiheit_3	**.36**	-.05	**.47**
unverfälschte Antwort_3	.01	**.40**	**.45**
Privatsphäre_1	**.32**	.16	**.34**
Eigenwert	4.73	3.25	2.67
erklärte Varianz (%)	20.58	14.13	11.59

Anhang 7.24c Faktorenanalyse der Akzept!-Skala zur Einstufung des Intelligenztests

n	194
KMO	.81
Kleinster KMO-Wert (Grenzwert .50)	.64
Bartlett–Test	χ^2 = 947, *df* = 120, *p* < .001
Determinante der Korrelationsmatrix (Grenzwert .00001)	$\|R\|$ = 0.00629
Test nach Haitovsky	χ^2_H = 1.18, *df* = 120, *p* > .05
Anzahl Iterationen	3
Erklärte Varianz	44.37



	Faktor 1	Faktor 2
Messqualität_2	**.78**	.01
Messqualität_3	**.74**	-.14
Augenscheinvalidität_2	**.68**	.16
Augenscheinvalidität_1	**.66**	.13
Messqualität_4	**.64**	-.10
Messqualität_1	**.61**	.25
Augenscheinvalidität_3	**.59**	-.05
Augenscheinvalidität_4	**.58**	.23
Kontrollierbarkeit_2	.00	**.80**
Kontrollierbarkeit_1	.05	**.76**
Kontrollierbarkeit_4	.12	**.66**
Belastungsfreiheit_1	.19	**.64**
Belastungsfreiheit_2	-.05	**.63**
Belastungsfreiheit_3	.20	**.61**
Kontrollierbarkeit_3	.00	**.57**
Belastungsfreiheit_4	-.03	**.33**
Eigenwert	3.64	3.46
erklärte Varianz (%)	22.75	21.61

Anhang 7.25a Korrelationen zwischen den Facetten des NEO-PI-R und den beiden Versionen des Leadership-Fragebogens

			Forced-Choice			likert-skaliert		
			DF	KF	VB	DF	KF	VB
			0.63	*0.83*	*0.54*	*0.83*	*0.91*	*0.79*
NEON1	Ängstlichkeit	*0.85*	-0.09	-0.38	-0.23	-0.16	-0.46	-0.37
NEON2	Reizbarkeit	*0.76*	0.02	-0.19	-0.11	-0.11	-0.17	-0.12
NEON3	Depression	*0.86*	-0.12	-0.36	-0.24	-0.27	-0.48	-0.37
NEON4	Befangenheit	*0.67*	-0.08	-0.47	-0.24	-0.20	-0.53	-0.40
NEON5	Impulsivität	*0.66*	-0.06	0.01	-0.04	-0.08	0.01	0.03
NEON6	Verletzlichkeit	*0.83*	-0.18	-0.26	-0.26	-0.22	-0.34	-0.37
NEON	Neurotizismus	*0.94*	-0.11	-0.35	-0.24	-0.22	-0.42	-0.34
NEOE1	Herzlichkeit	*0.76*	-0.15	0.54	0.12	-0.01	0.51	0.48
NEOE2	Geselligkeit	*0.81*	-0.12	0.59	0.00	-0.10	0.60	0.31
NEOE3	Durchsetzungsfähigkeit	*0.85*	0.20	0.45	0.34	0.32	0.57	0.47
NEOE4	Aktivität	*0.66*	0.13	0.20	0.09	0.11	0.27	0.20
NEOE5	Erlebnishunger	*0.66*	0.09	0.31	-0.03	-0.02	0.36	0.09
NEOE6	Frohsinn	*0.78*	-0.11	0.40	0.17	0.08	0.41	0.41
NEOE	Extraversion	*0.90*	0.02	0.62	0.18	0.10	0.68	0.48
NEOO1	Phantasie	*0.79*	-0.03	-0.04	-0.15	-0.04	-0.03	0.04
NEOO2	Ästhetik	*0.85*	-0.02	-0.12	-0.11	-0.01	-0.18	-0.03
NEOO3	Gefühle	*0.78*	-0.05	-0.05	0.08	0.04	-0.12	0.12
NEOO4	Handlungen	*0.58*	-0.07	0.33	0.20	-0.05	0.32	0.23
NEOO5	Ideen	*0.71*	0.11	0.07	0.06	0.06	0.07	0.06
NEOO6	Werte	*0.27*	-0.10	0.01	0.06	-0.02	0.09	0.10
NEOO	Offenheit	*0.87*	-0.03	0.02	0.01	0.00	0.00	0.11
NEOA1	Vertrauen	*0.75*	-0.04	0.23	0.20	-0.07	0.24	0.32
NEOA2	Freimütigkeit	*0.58*	-0.27	-0.25	-0.09	-0.35	-0.28	-0.06
NEOA3	Altruismus	*0.66*	-0.22	0.09	0.06	-0.18	0.03	0.29
NEOA4	Entgegenkommen	*0.55*	-0.34	-0.13	-0.24	-0.45	-0.20	-0.13
NEOA5	Bescheidenheit	*0.78*	-0.20	-0.29	-0.06	-0.18	-0.29	-0.17
NEOA6	Gutherzigkeit	*0.69*	-0.34	0.01	0.03	-0.28	-0.08	0.11
NEOA	Verträglichkeit	*0.85*	-0.35	-0.09	-0.02	-0.38	-0.15	0.08
NEOC1	Kompetenz	*0.73*	0.23	0.06	0.20	0.28	0.10	0.37
NEOC2	Ordnungsliebe	*0.78*	0.27	-0.20	0.06	0.29	-0.18	0.09
NEOC3	Pflichtbewusstsein	*0.65*	0.16	0.00	0.15	0.13	-0.02	0.19
NEOC4	Leistungsstreben	*0.74*	0.20	0.02	0.32	0.25	0.00	0.21
NEOC5	Selbstdisziplin	*0.82*	0.18	-0.08	0.14	0.21	-0.04	0.12
NEOC6	Besonnenheit	*0.76*	0.14	-0.22	0.02	0.13	-0.25	0.05
NEOC	Gewissenhaftigkeit	*0.93*	0.25	-0.10	0.18	0.28	-0.10	0.21

Anmerkung. *N* = 100. In kursiv sind die Skalenreliabilitäten nach Cronbach Alpha aufgeführt.

Anhang 7.25b Doppelt miderungskorrigierte Korrelationen zwischen dem NEO-PI-R und dem Leadership-Fragebogen

			Forced-Choice			likert-skaliert		
			DF	KF	VB	DF	KF	VB
			0.63	*0.83*	*0.54*	*0.83*	*0.91*	*0.79*
NEON1	Ängstlichkeit	*0.85*	-0.13	-0.45	-0.33	-0.19	-0.52	-0.45
NEON2	Reizbarkeit	*0.76*	0.03	-0.24	-0.17	-0.13	-0.20	-0.16
NEON3	Depression	*0.86*	-0.16	-0.43	-0.36	-0.32	-0.54	-0.45
NEON4	Befangenheit	*0.67*	-0.13	-0.64	-0.40	-0.27	-0.67	-0.55
NEON5	Impulsivität	*0.66*	-0.09	0.01	-0.07	-0.11	0.01	0.04
NEON6	Verletzlichkeit	*0.83*	-0.25	-0.31	-0.38	-0.27	-0.39	-0.45
NEON	Neurotizismus	*0.94*	-0.14	-0.39	-0.33	-0.25	-0.45	-0.40
NEOE1	Herzlichkeit	*0.76*	-0.21	0.68	0.18	-0.02	0.61	0.63
NEOE2	Geselligkeit	*0.81*	-0.17	0.73	0.00	-0.12	0.69	0.39
NEOE3	Durchsetzungsfähigkeit	*0.85*	0.28	0.54	0.51	0.39	0.65	0.58
NEOE4	Aktivität	*0.66*	0.20	0.27	0.15	0.15	0.34	0.27
NEOE5	Erlebnishunger	*0.66*	0.15	0.41	-0.04	-0.03	0.47	0.12
NEOE6	Frohsinn	*0.78*	-0.15	0.50	0.26	0.10	0.48	0.52
NEOE	Extraversion	*0.90*	0.03	0.72	0.26	0.12	0.75	0.58
NEOO1	Phantasie	*0.79*	-0.04	-0.05	-0.23	-0.04	-0.03	0.05
NEOO2	Ästhetik	*0.85*	-0.03	-0.14	-0.16	-0.02	-0.21	-0.04
NEOO3	Gefühle	*0.78*	-0.08	-0.06	0.12	0.05	-0.14	0.15
NEOO4	Handlungen	*0.58*	-0.11	0.47	0.36	-0.08	0.43	0.35
NEOO5	Ideen	*0.71*	0.16	0.10	0.10	0.08	0.09	0.09
NEOO6	Werte	*0.27*	-0.25	0.01	0.16	-0.04	0.18	0.22
NEOO	Offenheit	*0.87*	-0.05	0.03	0.01	0.00	0.00	0.14
NEOA1	Vertrauen	*0.75*	-0.06	0.30	0.31	-0.08	0.29	0.42
NEOA2	Freimütigkeit	*0.58*	-0.46	-0.36	-0.16	-0.50	-0.39	-0.09
NEOA3	Altruismus	*0.66*	-0.34	0.12	0.10	-0.24	0.03	0.40
NEOA4	Entgegenkommen	*0.55*	-0.57	-0.19	-0.43	-0.67	-0.28	-0.20
NEOA5	Bescheidenheit	*0.78*	-0.28	-0.36	-0.09	-0.22	-0.35	-0.22
NEOA6	Gutherzigkeit	*0.69*	-0.51	0.01	0.04	-0.36	-0.10	0.15
NEOA	Verträglichkeit	*0.85*	-0.48	-0.11	-0.03	-0.45	-0.17	0.10
NEOC1	Kompetenz	*0.73*	0.33	0.08	0.32	0.36	0.12	0.48
NEOC2	Ordnungsliebe	*0.78*	0.38	-0.24	0.09	0.36	-0.21	0.11
NEOC3	Pflichtbewusstsein	*0.65*	0.25	0.00	0.24	0.18	-0.02	0.26
NEOC4	Leistungsstreben	*0.74*	0.30	0.02	0.50	0.32	0.00	0.27
NEOC5	Selbstdisziplin	*0.82*	0.25	-0.09	0.21	0.26	-0.04	0.15
NEOC6	Besonnenheit	*0.76*	0.20	-0.27	0.02	0.16	-0.30	0.06
NEOC	Gewissenhaftigkeit	*0.93*	0.33	-0.12	0.26	0.32	-0.10	0.24

Anmerkung. N = 100. In kursiv sind die Skalenreliabilitäten nach Cronbach Alpha aufgeführt.

8. Zusammenfassung und Diskussion

8.1 Zusammenfassende Darstellung der Vorgehensweise bei der Testkonstruktion und der wichtigsten Ergebnisse

Ausgangspunkt der hier beschriebenen Entwicklung des Leadership-Fragebogens war die 1999 begonnene Neukonzeption der Aushebung der Schweizer Armee. Unter anderem gaben damals die Armeeplaner vor, dass in der neu als Rekrutierung bezeichneten Musterung der Stellungspflichtigen auch eine erste Erfassung des Kaderpotenzials stattfinden soll. Dabei soll auch ein Instrument zum Einsatz kommen, welches einige dafür relevante Persönlichkeitseigenschaften – zusammengefasst unter dem Begriff soziale Kompetenzen – erhebt. Ausgehend vom Auftrag der Leitung des Projektes Rekrutierung A XXI, verfolgte ich bei der Entwicklung dieses – von mir als Leadership-Fragebogen bezeichneten – Testverfahrens folgende drei Ziele:

1. Der als Situational Judgment Test (SJT) konzipierte Leadership-Fragebogen erfasst a priori definierte Persönlichkeitsdimensionen.
2. Die Fragebogenentwicklung basiert auf einem Konstruktionsrational.
3. Die Stellungspflichtigen akzeptieren den Leadership-Fragebogen gut.

Zu Beginn der Testentwicklungsarbeiten stand die Erstellung eines Anforderungsprofils für unteres Milizkader der Schweizer Armee. Nach dem Studium der dafür relevanten Literatur aus dem militärischen Umfeld befragten wir 22 Berufsmilitärs der Schweizer Armee anhand der *Critical Incident Technique*, was zu einer umfassenden Liste mit erfolgsrelevanten Verhaltensweisen führte. Mit inhaltsanalytischen Verfahren erstellten wir daraus ein aus 14 Anforderungsdimensionen bestehendes Kategoriensystem. Je acht Verhaltensweisen pro Anforderungsdimension liessen wir hinsichtlich deren Wichtigkeit für die Arbeit des unteren Milizkaders von 60 Berufsmilitärs einstufen, um daraus das Basisanforderungsprofil zu erstellen. Auf dieser Grundlage definierten die Mitglieder der Arbeitsgruppe Kaderselektion der Schweizer Armee das definitive Anforderungsprofil für unteres Kader der Schweizer Armee, welches folgende Dimensionen beinhaltet: Auffassungsgabe, Organisationsfähigkeit, Engagement und Eigeninitiative, Belastbarkeit und Beharrlichkeit, Gewissenhaftigkeit, Durchsetzungsfähigkeit, Führungs- und Verantwortungsbereitschaft, Konfliktverhalten, Kommunikation, Kontaktverhalten. Der neu zu entwickelnde Leadership-Fragebogen deckt daraus die Anforderungsdimensionen Durchsetzungsfähigkeit, Kontaktfähigkeit und Verantwortungsbewusstsein ab – also alles Persönlichkeits-

eigenschaften, welche im Kontakt mit den Unterstellten gefordert sind. In einer zusätzlichen Datenerhebung liess ich Gruppen– und Zugführer nochmals die Wichtigkeit der 112 Verhaltensweisen für ihre Funktion einstufen. Die Auswertung der 55 respektive 51 berücksichtigten Fragebogen ergab, dass sich die Rangreihenfolge der 14 Anforderungsdimensionen zwischen diesen beiden Gruppen nur bei der Teamfähigkeit signifikant unterscheiden, wobei die Gruppenführer diese wichtiger einstufen (Rangplatz 8) als die Zugführer (Rangplatz 13). Die Einstufungen der Gruppen- und Zugführer zusammengenommen unterscheiden sich jedoch deutlich – in den Anforderungsdimensionen Verantwortungsübernahme, Gewissenhaftigkeit & Loyalität, physische und psychische Belastbarkeit, Fürsorglichkeit / Einfühlungsvermögen und Teamfähigkeit – von denjenigen der Berufsmilitärs. Grosse Unterschiede zeigten sich auch beim Vergleich der Einstufungen bei den verschiedenen Lehrverbänden, jedoch nur bei den Gruppenführern.

Zur Generierung der Item-Stämme des als Situational Judgment Test konzipierten Leadership-Fragebogens setzten wir den Act Frequency Approach (AFA; Buss & Craik, 1980, 1984) ein und orientierten uns dabei an der von Krüger und Amelang (1995) beschriebenen Vorgehensweise. Wir liessen 38 Schülerinnen und Schüler im Alter von 18 bis 20 Jahren insgesamt ungefähr 240 Acts zu den drei Anforderungsdimensionen beschreiben, wobei wir bei der Überarbeitung der Acts einige davon auf Grund hoher Ähnlichkeit ausschlossen. Zur Bestimmung der Prototypizität der verbleibenden 149 Acts liessen wir diese von knapp 40 Probanden einstufen, wobei die durchschnittliche Prototypizität bei den Acts zur Dimension Durchsetzungsfähigkeit leicht tiefer ausfiel als diejenigen der beiden anderen Dimensionen. Als Ausgangsmaterial für die Entwicklung der Situational Judgment Test-Items wählten wir die prototypischsten Acts aus und formulierten einen eindeutigen, auf die Erfahrungsrealität von 19jährigen abgestimmten Situationskontext dazu.

Als Konstruktionsrational für die Formulierung der vier Verhaltensweisen pro Item-Stamm haben wir das Wertequadrat (Helwig, 1948, 1967; siehe auch Gloor, 1993; Schulz von Thun, 1989; Westermann, 2007) verwendet. In einem ersten Schritt suchten wir für die vier Wertequadranten pro zu erfassende Persönlichkeitsdimension nach passenden Labels und nahmen in einem zweiten Schritt eine genaue Definition jedes einzelnen Wertequadranten vor. Auf dieser Grundlage konstruierten wir insgesamt 81 Items, welche wir in vier Pretests insgesamt 643 Rekruten aus fünf verschiedenen Rekrutenschulen zur Bearbeitung vorlegten. Daraus bildeten wir die erste definitive Version des Leadership-Fragebogens, welche sich aus je 13 Items pro Dimension zusammensetzte, wobei jedes Item aus einer Situationsbeschreibung, einer die Situation illustrierenden

Fotografie und den vier anhand des Wertequadrates erstellten Verhaltensalternativen besteht. Die Testbearbeiter erhalten die Instruktion, diejenige Verhaltensalternative auszuwählen, welche sie in der geschilderten Situation am ehesten zeigen würden. Die Bewertung der Antworten erfolgt anhand des Ausprägungsgrades der Verhaltensalternative innerhalb der gemessenen Dimension: Bei der Dimension Kontaktfähigkeit ergibt so die Verhaltensalternative, welche dem Wertequadranten 1 „Menschenscheu" zugeordnet ist, einen Punkt, diejenige der „Zurückhaltung" zwei Punkte, „Kontaktfähigkeit" drei Punkte und die Verhaltensalternative zum Wertquadranten 4 „Distanzlosigkeit" wird mit vier Punkten bewertet.

Die Analyse dieser ersten Version anhand eines 7'871 Stellungspflichtige[1] umfassenden Datensatzes ergab folgende Reliabilitäten (in Klammern habe ich die Reliabilitäten der auf zehn Items gekürzten Skalen aufgeführt): Durchsetzungsfähigkeit: α = .56 (.53), Kontaktfähigkeit: α = .84 (.83), Verantwortungsbewusstsein: α = .74 (.71), womit die Skala Durchsetzungsfähigkeit die Anforderungen an die Reliabilität einer im Selektionskontext eingesetzten Skala eindeutig nicht erfüllt (Evers, 2001; Lindley, Bartram & Kennedy, 2008). Die Skalen Kontaktfähigkeit und Verantwortungsbewusstsein korrelieren mit r = .56 (.53) und stehen in keinem Zusammenhang zur Skala Durchsetzungsfähigkeit. Die explorative Faktorenanalyse ergab drei klar identifizierbare Faktoren, welche jedoch nur gerade 26.46% (30.01%) der Varianz erklären. Anhand eines Datensatzes mit den Angaben von 19'801 Stellungspflichtigen konnte ich die Kennwerte des auf 30 Items gekürzten Fragebogens bestätigen. Zudem konnte ich mit einer an das Konzept der Distraktorenanalyse angelehnten Methode nachweisen, dass die Operationalisierung der Wertequadrate bei allen Items gelungen ist.

In einer zusätzlichen Studie habe ich die Auswirkungen unterschiedlicher Scoring-Arten auf die Reliabilität und die Faktorenstruktur der Leadership-Skalen untersucht. Es zeigte sich jedoch, dass verschiedene, vom ursprünglich gewählten Scoring abweichende Gewichtungen der vier Antwortalternativen zu schlechteren Reliabilitäten führen. Weiter untersuchte ich die Auswirkungen auf die Reliabilität bei der Einführung eines likert-skalierten Antwortformates, bei welchem der Testbearbeiter bei jeder Verhaltensalternative auf einer vierstufigen Skala angeben kann, wie zutreffend diese für ihn ist. Die anhand der Datensätze von 1'017 Stellungspflichtigen berechneten Reliabilitäten liegen bei allen drei Dimensionen über .80 und lassen sogar – nur den Aspekt der Homogenität der Skalen

[1] Eine Übersicht über alle in dieser Arbeit verwendeten Stichproben ist im Anhang in Tabelle 8.1 dargestellt.

berücksichtigend – eine Kürzung des Fragebogens auf sechs Items pro Skala zu. Die explorative Faktorenanalyse ergibt fünf Faktoren, welche insgesamt 27.99% der Varianz aufklären. Die Analyse der Itemladungen zeigt, dass der Einsatz des likert-skalierten Antwortformates zu einem Methodenfaktor führt, auf welchen mehrheitlich die Verhaltensalternativen zum Wertequadranten 2 laden. Der fünfte Faktor bildet sich aus Verhaltensalternativen des ersten Wertequadranten der Dimension Verantwortungsbewusstsein und denjenigen des vierten Wertequadranten der Dimension Durchsetzungsfähigkeit. Eine forcierte Dreifaktorenlösung unter Ausschluss des Wertequadranten 2 bildet die drei Dimensionen des Leadership-Fragebogens relativ gut ab, jedoch bleibt die Varianzaufklärung weiterhin ungenügend. Die gewünschte Faktorenstruktur lässt sich erst erzielen, nachdem ich die drei Leadership-Skalen auf je sechs Item-Stämme reduziert habe und die Faktorenanalyse nicht mehr über die 72 respektive 54 einzelnen Verhaltensalternativen, sondern über die 18 Items rechne. Sowohl bei der Berechnung mit vier Wertequadranten wie auch bei derjenigen mit drei klären die drei Faktoren 50% der Varianz auf und es treten keine Fehlladungen mehr auf.

Die anhand eines Datensatzes von 100 Studierenden berechneten Korrelationen zwischen der Forced-Choice- und der likert-skalierten Version des Leadership-Fragebogens liegen zwischen .64 und .88. Eine doppelte Minderungskorrektur führt zu einem perfekten Zusammenhang der jeweiligen Dimensionen der beiden Versionen. Dies ist ein Hinweis darauf, dass das Antwortformat die Konstruktvalidität nicht beeinflusst.

Anhand einer studentischen Stichprobe mit 35 männlichen Probanden im Alter zwischen 18 und 24 Jahren überprüfte ich den Bekanntheitsgrad der im Leadership-Fragebogen beschriebenen Situationen. Dieser fiel erwartungsgemäss sehr unterschiedlich aus und reichte von einer Situation, welche nur einer der Probanden schon einmal erlebt hatte, bis zu einer Situation, welche bis auf einen Probanden alle schon einmal erlebt hatten. Zusätzlich stellte ich noch die Frage, ob man sich in die geschilderte Situation hineinversetzen kann. Hier fielen die Zustimmungsraten deutlich höher aus und liegen zwischen 71% und 100%. Weiter ging ich der Frage nach, ob der Bekanntheitsgrad eines Items einen Einfluss auf dessen Messgenauigkeit hat. Die durchschnittliche Korrelation zwischen der Frage, ob man die Situation schon einmal erlebt hat und der Trennschärfe des Items beträgt .33 und .44 bei der Frage, ob man sich in die beschriebene Situation hineinversetzen kann. Dies ist ein Hinweis darauf, dass die Probanden ein konsistenteres Antwortmuster zeigen, wenn sie ihr Verhalten in Situationen einschätzen müssen, welche sie konkret erlebt haben, als wenn sie hypothetische Aussagen zu ihrem Verhalten machen müssen. Zudem zeigt dieses Ergebnis auf, dass man bei einer auf dem Act Frequency Approach basierenden

Testentwicklung nicht nur die Prototypizität der Situationen einstufen lassen muss, sondern auch deren Bekanntheitsgrad.

In einem aufwändigen Verfahren haben wir die Akzeptanz des Leadership-Fragebogens untersucht. In einer ersten Studie setzten wir dazu einen Fragebogen mit 15 Items zu den Aspekten Layout, Verständlichkeit, Erleben, Augenscheinvalidität, Alltäglichkeit und Privatsphäre ein, um die Akzeptanzeinschätzung des Leadership-Fragebogens mit denjenigen von zwei Persönlichkeits-Fragebogen mit likert-skalierten und Forced-Choice-Antwortformat zu vergleichen. Insgesamt schätzten 131 Schüler, Rekruten und Unteroffiziers-Schüler diese drei Testverfahren ein, wobei der Leadership-Fragebogen – mit Ausnahme beim Aspekt Augenscheinvalidität – die höchsten Einstufungen erhielt und der Forced-Choice-Fragebogen die tiefsten. Die grössten Unterschiede haben sich bei der Einstufung des Layouts ergeben. In einer zweiten Studie habe ich den Effekt der die Situation illustrierenden Fotografie auf die Akzeptanzeinschätzung des Leadership-Fragebogens untersucht, indem ich als Vergleich dazu eine Version des Leadership-Fragebogens ohne Fotografie und einen likert-skalierten Persönlichkeits-Fragebogen einstufen liess. Insgesamt haben 717 angehende Unteroffiziere die Akzeptanz der drei Fragebogen mit einer leicht abgeänderten Version des in der ersten Studie eingesetzten Akzeptanz-Fragebogens eingeschätzt. Die beiden Versionen des Leadership-Fragebogens unterscheiden sich nur bezüglich der Einschätzung des Layouts, welche bei der Fotografie-Version höher ist. Die beiden Versionen werden jedoch – mit Ausnahme in der Dimension Alltäglichkeit – in allen anderen Aspekten signifikant höher eingestuft als der likert-skalierte Persönlichkeits-Fragebogen. Eine schrittweise Regressionsanalyse der Akzeptanzvariablen auf das Erleben der Fragebogenbearbeitung ergab, dass die Variable Layout den höchsten Einfluss hat und 47.8% der Varianz erklärt.

Für die dritte Akzeptanz-Studie setzte ich den Akzept!-Fragebogen von Kersting (2008) ein, welcher die Aspekte Kontrollierbarkeit, Messqualität, Augenscheinvalidität, Wahrung der Privatsphäre, Intention zur unverfälschten Antwort, Antwortfreiheit und eine Gesamtbeurteilung erfasst und von mir mit den Aspekten „Spass an der Bearbeitung", „schöne Darstellung" und „sich in die Situation hineinversetzen können" ergänzt wurde. Ich liess je ungefähr 200 Stellungspflichtige die Akzeptanz von einem von drei in den Rekrutierungszentren eingesetzten Testverfahren einstufen: dem Leadership-Fragebogen, einem traditionellen Persönlichkeits-Fragebogen und einem Intelligenztest. Dabei fällt die Akzeptanzeinschätzung des Leadership-Fragebogens und des Persönlichkeits-Fragebogens ungefähr gleich aus, diejenige des Intelligenztests ist leicht tiefer. Bei der schrittweisen Regressionsanalyse der Akzeptanzvariablen auf die Gesamtbeurteilung des Leadership-Fragebogens stellt die Variable Darstellung mit 25% erklär-

ter Varianz den besten Prädiktor dar, gefolgt von der Augenscheinvalidität, dem Sich-Hineinversetzen-Können, der Messqualität und der Privatsphäre. Die Analyse des Akzept!-Fragebogens ergab, dass die Stellungspflichtigen nicht zwischen den Dimensionen Augenscheinvalidität und Messqualität differenzieren können.

Abschliessend habe ich anhand der Daten von 17'000 Stellungspflichtigen untersucht, ob ich das primäre Ziel der Testentwicklung erreicht habe und der Leadership-Fragebogen die darin operationalisierten Persönlichkeitsdimensionen erfasst. Die Analyse zeigt klar auf, dass die Dimensionen des Leadership-Fragebogens nicht mit verschiedenen Aspekten der Intelligenz korrelieren, wie dies sonst bei nach dem herkömmlichen Verfahren entwickelten SJTs der Fall ist. Mit den sechs Dimensionen (Leistungsmotivation, Belastbarkeit, Extraversion, Gewissenhaftigkeit, Entgegenkommen/Friedfertigkeit, Teamfähigkeit) des in den Rekrutierungszentren eingesetzten Persönlichkeits-Fragebogens korreliert die Dimension Durchsetzungsfähigkeit erwartungsgemäss negativ mit der Skala Entgegenkommen/Friedfertigkeit. Die beiden anderen Dimensionen zeigen mittlere bis hohe Korrelationen mit allen sechs Skalen, wobei Kontaktfähigkeit am höchsten mit Extraversion und Verantwortungsbewusstsein am höchsten mit Leistungsmotivation, Belastbarkeit und Gewissenhaftigkeit korreliert. Das Herauspartialisieren der Führungsmotivation, welche Anlass für eine bewusste Antwortverzerrung durch die Stellungspflichtigen sein könnte, verdeutlicht die gefundenen Korrelationsmuster zusätzlich.

Anhand des Datensatzes von 100 Studierenden berechnete ich den Zusammenhang der drei Skalen des Leadership-Fragebogens mit den Big Five, wozu ich den NEO-PI-R (Ostendorf & Angleitner, 2004). einsetzte. Durchsetzungsfähigkeit korreliert dabei wie erwartet negativ mit Verträglichkeit (r = -.48) und positiv mit Gewissenhaftigkeit (r = .33), Kontaktfähigkeit sehr stark mit Extraversion (r = .72) und negativ mit Neurotizismus (r = -.39). Die Skala Verantwortungsbewusstsein korreliert mit r = -.33 mit Neurotizismus und in der likert-skalierten Version noch mit r = .58 mit Extraversion. Anhand dieser Ergebnisse ist die Bestätigung erbracht, dass der Leadership-Fragebogen die a priori festgelegten Persönlichkeitseigenschaften erfasst und keine Aspekte kognitiver Leistungsfähigkeit miteinfliessen.

8.2 Diskussion der Testentwicklung

Entwicklung des Anforderungsprofils für unteres Milizkader

Für die Entwicklung des Anforderungsprofils für unteres Milizkader habe ich ein sehr aufwändiges Vorgehen gewählt, welches vier Informationsquellen nutzte: Die bestehende, militärspezifische Literatur; Berufsmilitärs zur Generierung erfolgsrelevanter Situationen; Berufsmilitärs, Gruppen- und Zugführer zur Einstufung der Wichtigkeit der extrahierten Verhaltensweisen und Vertreter der Stäbe aller Lehrverbänden der Armee zur Erstellung des definitiven Anforderungsprofils. Wie die Übersichtstabelle 6.9 in Kapitel 6.1 zeigt, hätte schon nur auf Grund der vorhandenen Literatur ein passables Anforderungsprofil erstellt werden können. Dies ist jedoch darauf zurückzuführen, dass zwei dieser Quellen – Annen (2000) und Hoenle (1996) – schon anhand empirischer Untersuchungen erstellte Anforderungskataloge enthalten.

Der grosse Vorteil der von mir durchgeführten Erhebung erfolgsrelevanter Verhaltensweisen mit der Critical Incident Technique (Flanagan, 1954) liegt darin, dass dieses Verfahren eine ausführliche Liste mit Führungstätigkeiten liefert, welche einerseits zur präzisen Definition der einzelnen Dimensionen des Anforderungsprofils dienen und anderseits eine Grundlage für die Konstruktion von Übungen im Rahmen der Kaderselektion darstellen. Zudem setzte ich die Listen mit den Verhaltensweisen dazu ein, um anhand einer Befragung von Berufsmilitärs eine Gewichtung der Anforderungsdimensionen zu erstellen. Ein bemerkenswertes Ergebnis dieser Datenerhebung ist die Einstufung der Wichtigkeit der Verhaltensweisen zur Dimension Teamfähigkeit, welche den letzten Platz der 14 eingestuften Dimensionen belegt. Anhand einer zusätzlichen Datenerhebung bei Gruppen- und Zugführern ging ich der Frage nach, ob dies nur die (Aussen-)Sicht der Berufsmilitärs widerspiegelt, welche eventuell mit Erfahrungen aus höheren Führungsstufen vermengt ist oder ob es tatsächlich so ist, dass die Teamfähigkeit für Gruppenführer nicht erfolgsentscheidend ist. Die Auswertung der Daten ergab, dass die Gruppenführer die Wichtigkeit der Teamfähigkeit signifikant höher einstufen als die Zugführer und die Berufsmilitärs. Somit scheint dieser Aspekt für die unterste Führungsstufe noch von mittlerer Wichtigkeit zu sein, was die Berücksichtigung dieser Dimension im Leadership-Fragebogen auch rechtfertigt. Allgemein zeigte sich jedoch ein deutlicher Unterschied zwischen den Einstufungen der Berufsmilitärs und denjenigen der Gruppen- und Zugführer, was im Gegensatz zu anderen publizierten Forschungsergebnissen steht, bei welchen sich die Einschätzungen eines kleinen Expertenteams nicht stark von denjenigen einer grossen Gruppe von Jobinhabern unterscheiden (z. B.

Maurer & Tross, 2000; Tannenbaum & Wesley, 1993). Das hier gefundene Ergebnis könnte darauf zurückzuführen sein, dass die Berufsmilitärs die Wichtigkeit der Anforderungsdimensionen danach einstufen, wie sie idealerweise sein sollte (Soll-Einstufung), wohingegen die Gruppen- und Zugführer diese auf Grund ihrer persönlichen Erfahrungen vornehmen (Ist-Einstufung).

Die Auswertung dieser Zusatzerhebung ergab noch einen zweiten, wichtigen Befund: Bei den Gruppenführern zeigten sich deutliche Unterschiede in den Einstufungen der Wichtigkeit der Anforderungsdimensionen zwischen den verschiedenen Lehrverbänden, bei den Zugführern hingegen nicht. Dies führt zu folgender Hypothese: Je höher die Gradstufe, desto einheitlicher werden die auszuführenden Tätigkeiten, was bewirkt, dass die Unterschiede in der Wichtigkeit der einzelnen Anforderungsdimensionen zwischen verschiedenen Truppengattungen kleiner werden. Dies hat zur Folge, dass zur Einstufung der Wichtigkeit der Anforderungsdimensionen auf der untersten Führungsstufe darauf zu achten ist, Gruppenführer aus allen Lehrverbänden in die Stichprobe aufzunehmen.

Das Datenscreening der Fragebogen der Gruppen- und Zugführer führte dazu, dass ich nur gut die Hälfte davon in die Auswertungen einbeziehen konnte, was sich jedoch mit Erfahrungswerten aus anderen Studien deckt (z. B. Green & Stutzman, 1986). Für zukünftige Datenerhebungen in den Rekrutenschulen ist deshalb zu prüfen, welches der in der Literatur vorgeschlagenen Verfahren – Wiederholung einzelner Items (Wilson, Harvey & Macy, 1990), Homogenität der Ratereinstufungen (Hughes & Prien, 1989), Einfügen von Kontrollverhaltensweisen (Green & Stutzman, 1986) oder Infrequenz-Index (Green & Veres, 1990) – sich am besten eignet, zuverlässig unseriös ausgefüllte Fragebogen zu eruieren. Um den Einfluss von Antworttendenzen zu minimieren, könnte anstelle der Wichtigkeitseinstufung jeder Verhaltensweise auch ein systematischer Paarvergleich eingesetzt werden, bei welchem der Jobinhaber bei jeweils zwei Verhaltensweisen aus unterschiedlichen Anforderungsdimensionen entscheiden muss, welche für seine Tätigkeit wichtiger ist (z. B. Opgenoorth, 1979).

Es stellt sich hier abschliessend noch die Frage, ob es zulässig und sinnvoll ist, ein anhand empirischer Daten erstelltes Anforderungsprofil schlussendlich noch von einem Fachgremium abändern zu lassen. Vom wissenschaftlichen Standpunkt aus betrachtet, führt dieses Vorgehen zu einer Verwässerung der auf empirischen Grundlagen erstellten Basis, da die subjektiven Meinungen der an einer solchen Arbeitsgruppe teilnehmenden Vertreter unter Umständen massgeblich und spürbar das Endprodukt beeinflussen. Von der Praxisseite her betrachtet, ist es jedoch unabdingbar, die Meinung wichtiger Vertreter der Organisation aufgenommen und integriert zu haben. Das Anforderungsprofil wird so

weniger als Fremdkörper wahrgenommen, sondern als eine unter Mitarbeit von Organisationsmitgliedern entwickelte Grundlage für die Kaderselektion. Vor allem im Hinblick auf die Einführung dieses Anforderungsprofils in der Armee ist eine grosse Akzeptanz wichtig. Hier lässt sich auch eine Brücke zur von Annen (2000) im Zusammenhang mit der Entwicklung und Einführung eines neuen Beurteilungssystems in der Schweizer Armee eingesetzten Aktionsforschung schlagen, in welcher der gesamte Prozess in enger Zusammenarbeit mit den Entscheidungsträgern vor Ort durchgeführt wird.

Das erstellte Anforderungsprofil bildet nun die Basis für den gesamten Selektionsprozess und alle darin enthaltenen Verfahren. Wenn es nur um die isolierte Entwicklung des Leadership-Fragebogens gegangen wäre, hätte definitiv nicht ein derart aufwändiges Analyseverfahren durchgeführt werden müssen, da das Instrument ja nur einen Ausschnitt aus den führungsrelevanten Eigenschaften abbilden soll. Es hätte in diesem Fall wohl ausgereicht, die Literatur zu konsultieren und mit wenigen, sorgfältig ausgewählten Berufsmilitärs ein halbstrukturiertes Interview durchzuführen.

Der Act Frequency Approach als Methode zur Generierung der Item-Stämme

Auch wenn der Act Frequency Approach (AFA) einige methodische Schwächen aufweist und zurecht als Königsweg zur Erforschung der Persönlichkeit umstritten ist (Angleitner & Demtröder, 1988; Block, 1989; Larsen & Buss, 2001; Moser, 1989), eignet er sich hervorragend für die Sammlung des Ausgangsmaterials für die Konstruktion der Item-Stämme zu den einzelnen Persönlichkeitsdimensionen. Die Anweisungen zum Vorgehen bei der Generierung von Acts waren für unsere Probanden – Schülerinnen und Schüler im Alter zwischen 18 und 20 Jahren – gut verständlich und sie waren grundsätzlich motiviert, während des regulären Unterrichts eine Stunde dafür zu investieren.

Die Bestimmung der durchschnittlichen Prototypizität der einzelnen Acts pro Dimension ergab bei der Durchsetzungsfähigkeit (M = 2.83) einen tieferen Wert als bei der Kontaktfähigkeit (M = 3.05) und dem Verantwortungsbewusstsein (M = 3.00), wobei dieser Unterschied nicht signifikant ist. Dieses Ergebnis gibt jedoch zumindest einen Hinweis darauf, dass es von der vorgegebenen Persönlichkeitseigenschaft abhängt, wie viele hochprototypische Situationen eine bestimmte Probandenpopulation dazu beschreiben kann. Eine tiefe durchschnittliche Prototypizität könnte dadurch bedingt sein, dass diese Dimension repräsentierendes Verhalten im Alltag allgemein nicht so häufig gezeigt wird oder dass die Dimension zu wenig klar definiert wurde, was dazu führt, dass nicht alle Probanden genau dieselbe Vorstellung von für diese Eigenschaft typischen Verhaltens-

weisen haben. Dass nur wenig hochprototypische Verhaltensweisen generiert werden, könnte jedoch auch darauf zurückzuführen sein, dass der Erfahrungsschatz mit Situationen, in welchen die jeweilige Persönlichkeitseigenschaft gezeigt wird, für die befragte Probandenpopulation gering ist: Hat man bisher nur wenig Erfahrungen mit der entsprechenden Persönlichkeitseigenschaft sammeln können, ist es nahe liegend, dass sich darunter auch weniger hochprototypische Verhaltensweisen befinden. In diesem Fall müsste eine andere Probandengruppe gesucht werden, welche über mehr Erfahrung mit Situationen verfügt, in welcher typischerweise der entsprechenden Persönlichkeitsdimension zugeordnetes Verhalten gezeigt wird.

Nachfolgende Liste führt die Punkte auf, welche es – auf Grund der von mir gemachten Erfahrungen beim Einsatz des AFAs – im Rahmen der Entwicklung eines SJTs zu beachten gilt:

- Damit sich die Acts möglichst gut mit dem Erfahrungsschatz der Population decken, für welche der SJT entwickelt wird, sind für deren Generierung Probanden zu wählen, welche aus dieser Population stammen oder mit ihr grosse Ähnlichkeit aufweisen. Dabei muss auch darauf geachtet werden, dass die Probanden in ihrem bisherigen Leben schon viele Situationen erlebt haben, welche für die zu erfassende Dimension prototypisch sind.
- Da nur diejenigen Acts mit einer hohen Prototypizitätseinstufung als Ausgangsmaterial für die Entwicklung eines Item-Stamms dienen und sich einige der geschilderten Situationen auf Grund ihres Inhaltes nicht dafür eignen, muss eine grosse Anzahl möglichst vielfältiger Acts generiert werden. Pro Persönlichkeitsdimension sollten mindestens 200 verschiedene Acts vorliegen. Diese Zahl ist deutlich tiefer, als die Zahlen, welche in der Literatur zur Entwicklung von SJTs genannt werden (Latham & Wexley, 1982; McHenry & Schmitt, 1994; Motowidlo, Dunnette & Carter, 1990; Muck, Höft, Hell & Schuler, 2006). Dabei ist jedoch zu beachten, dass sich meine Angabe auf eine Persönlichkeitsdimension bezieht, die Angaben in oben erwähnter Literatur jedoch auf einen SJT, welcher die Verhaltensweisen in einem Tätigkeitsgebiet abdecken soll.
- Damit sich die Dimensionen im SJT möglichst gut voneinander unterscheiden, ist die von Angleitner und Demtröder (1988) vorgeschlagene Mehrfachsortierung durchzuführen, bei welcher die Probanden die Prototypizität jedes Acts zu allen zu erfassenden Persönlichkeitsdimensionen einschätzen. So können für die weitere Testentwicklung diejenigen

Acts verwendet werden, welche hochprototypisch für die intendierte Persönlichkeitsdimension und niedrigprototypisch für die restlichen Dimensionen sind.

- Um möglichst augenscheinvalide Item-Stämme entwickeln zu können, ist zusätzlich zum Prototypenrating auch noch ein Rating des Bekanntheitsgrades oder der Auftretenshäufigkeit durchzuführen. Damit lassen sich diejenigen Acts bestimmen, welche nur wenige der Probanden schon einmal selbst erlebt haben, um diese dann entweder umzuformulieren oder auszuschliessen.

Die Vorteile des Einsatzes des AFA im Rahmen der Testkonstruktion sind offensichtlich: Im Vergleich zu herkömmlichen Persönlichkeits-Fragebogen-Items sind anhand des AFAs entwickelte Items deutlich näher am Alltagserleben der zukünftigen Testbearbeiter, was dazu führen sollte, dass sie den Fragebogen besser akzeptieren. Zudem ist der Testentwickler bei der Itemgenerierung weniger auf seinen „sixth sense" (Osterlind, 1998, S. 2) angewiesen und kann sich bei seiner Aufgabe auf umfangreiches Ausgangsmaterial stützen. Vor allem bei theoretisch schwer fassbaren Konstrukten bildet dieses Material eine gute Grundlage, das Konstrukt besser verstehen zu können (z. B. Cinite, Duxbury & Higgins, 2009).

Das Wertequadrat als Konstruktionsrational zur Entwicklung der Verhaltensalternativen

Ich habe mir für diese Testentwicklung selbst die Auflage gegeben, dass sie auf einem Konstruktionsrational basiert. Damit wollte ich erreichen, dass die zu den Item-Stämmen entwickelten Verhaltensalternativen über alle Items einer Persönlichkeitsdimension hinweg vergleichbar sind. Dies bedeutet jedoch, dass ich mich vom üblichen Vorgehen bei der Entwicklung eines SJTs abgewendet habe, indem ich darauf verzichtet habe, die Item-Stämme Experten vorzulegen, welche dann mehr oder weniger effektive Verhaltensalternativen dazu formuliert hätten.

Wie sich im Verlauf der Testentwicklung zeigte, eignet sich das Wertequadrat (Helwig, 1948, 1967) für diesen Zweck hervorragend und die Datenanalyse zeigt klar auf, dass es uns auch gelungen ist, dessen Logik in der Forced-Choice-Version des Leadership-Fragebogens umzusetzen. Damit ersparte ich mir zwei Expertenbefragungen: Die erste, um die Verhaltensalternativen zu generieren und die zweite für die Einstufung deren Effektivität.

Einzig wenn bei der likert-skalierten Version des Leadership-Fragebogens jede Verhaltensalternative als eigenständiges Item in die Berechnungen aufgenommen wird, ergibt die explorative Faktorenanalyse nicht die erwarteten drei

Faktoren sondern deren fünf. Rechne ich hingegen die Faktorenanalyse über die zu einem Score summierten Wertequadranten pro Item-Stamm, so ergeben sich wieder die drei Leadership-Dimensionen. Die beiden zusätzlichen Faktoren der likert-skalierten Version lassen sich auf die Verwendung des Wertequadrates als Konstruktionsrational zurückführen: Der eine setzt sich hauptsächlich aus Verhaltensalternativen des zweiten Wertequadranten zusammen und der andere aus dem Wertequadranten 4 der Skala Durchsetzungsfähigkeit und dem Wertequadranten 1 der Skala Verantwortungsbewusstsein, also aus zwei Übertreibungen der beiden Skalen. Dabei könnten Antworttendenzen (*Response Styles* resp. *Bias*; z. B. Mummendey & Grau, 2008; Nunnally & Bernstein, 1994) für die Ausbildung dieser Faktoren verantwortlich sein: Der erste Methodenfaktor könnte sich dadurch herausbilden, dass sich viele der Probanden als ausgeglichen und durchschnittlich (Tugenden), ohne Ecken und Kanten (Übertreibungen) dargestellt haben. Dies führt dazu, dass sie unabhängig von der geschilderten Situation und der damit erfassten Persönlichkeitseigenschaft eine hohe Ausprägung bei der einen Tugend wählen und bei den Übertreibungen dementsprechend tiefe Ausprägungen. Auch der zweite Methodenfaktor liesse sich so erklären: Die beiden in diesem Faktor zusammengefassten Wertequadranten sind Rücksichtslosigkeit und die Verantwortungslosigkeit. Auch wenn die Probanden diese Labels nicht erfahren haben, so sind es wahrscheinlich die beiden Übertreibungen, welche – vielleicht vor allem für Studierende – wohl am wenigsten sozial erwünscht sind. Die anderen vier Übertreibungen – Selbstverleugnung, Menschenscheu, Distanzlosigkeit und Bevormundung – scheinen da weniger auf Ablehnung zu stossen.

8.3 Diskussion der Testüberprüfung

Reliabilität und faktorielle Validität des Leadership-Fragebogens

Die verschiedenen Reliabilitätsanalysen der Forced-Choice-Version des Leadership-Fragebogens zeigten alle dasselbe Bild: Die Reliabilitäten der Skalen Kontaktfähigkeit und Verantwortungsbewusstsein liegen zwischen α = .70 und .84, diejenige der Skala Durchsetzungsfähigkeit um α = .55, was die Minimalgrenze von α = .70 eindeutig verfehlt (Evers, 2001; Lindley, Bartram & Kennedy, 2008). Erfreulicher fällt die Überprüfung der Faktoren aus, indem die Drei-Faktoren-Struktur bestätig wird und alle Items auf den richtigen Faktor laden. Zudem entsprechen die Antwortverteilungen bei jedem Item der Logik des Wertequadrates.

Die Einführung des likert-skalierten Antwortformates führt wie erwartet zu einer deutlichen Erhöhung der Skalenreliabilitäten von über .80. Als Nachteil handelt man sich jedoch – zumindest wenn man alle Verhaltensalternativen einzeln einbezieht – eine inhaltlich nicht mehr erklärbare Faktorenstruktur ein. Dieses Problem lässt sich entschärfen, indem man von den zu den beiden Tugenden entwickelten Verhaltensalternativen eine ausschliesst. Damit ist das Wertequadrat zwar nicht mehr vollständig umgesetzt, dafür gewinnt der Test an Validität und die Bearbeitungsdauer sinkt. Zusätzlich wird es auch deutlich einfacher, die Verhaltensalternativen für die zwei Übertreibungen und nur eine Tugend zu verfassen, da es bei der Testentwicklung die grösste Herausforderung darstellte, die Verhaltensweisen zu den beiden Tugenden so zu formulieren, dass sie sich auch deutlich voneinander unterscheiden.

Somit schlage ich die Entwicklung einer neuen Version des Leadership-Fragebogens vor, welche pro Item-Stamm nur noch drei Verhaltensalternativen umfasst, welche jedoch je einzeln anhand eines likert-skalierten Antwortformates einzustufen sind. Damit sollte es möglich sein, einen Test zu erstellen, dessen Bearbeitung ungefähr zehn bis 15 Minuten dauert und welcher die drei Dimensionen reliabel erfasst.

Zum Bekanntheitsgrad der im Leadership-Fragebogen geschilderten Situationen

Die Überprüfung des Bekanntheitsgrades der in den Leadership-Items enthaltenen Situationen („Haben Sie eine solche oder ähnliche Situation, wie sie in der Ausgangslage geschildert ist, schon einmal erlebt?") ergab grosse Unterschiede zwischen den einzelnen Dimensionen: So haben im Schnitt 64% der Probanden die in der Dimension Kontaktfähigkeit geschilderten Situationen schon einmal selbst erlebt, jedoch nur gerade 32% diejenigen der Dimension Verantwortungsbewusstsein. Bei der Frage danach, ob man sich in diese Situation hineinversetzen könne, liegt die durchschnittliche Zustimmungsrate bei 89% bis 96%. Die – mit Ausnahme bei der Dimension Durchsetzungsfähigkeit – im mittleren Bereich liegenden Korrelationen zwischen dem Bekanntheitsgrad und der Trennschärfe eines Items zeigen auf, dass es für die psychometrische Güte eines Items nicht unbedeutend ist, wie bekannt die darin geschilderte Situation ist.

Dieses Ergebnis besagt somit nichts anderes, als dass Personen ein konsistenteres Bild ihres Verhaltens abgeben, wenn sie von realem Verhalten berichten, als wenn sie sich zu hypothetischem Verhalten äussern. Dies scheint naheliegend zu sein, im Zusammenhang mit Selektionsinterviews wird jedoch noch heute darüber debattiert, ob das situative Interview mit Fragen nach hypothetischen Verhaltensweisen in künftigen Situationen (Latham, Saari, Purcell &

Campion, 1980) dem Verhaltensbeschreibungs-Interview mit Fragen nach früherem anforderungsrelevantem Verhalten (*Behavior Description Interview*; Janz, 1982, 1989) überlegen ist (z. B. Jetter, 2008; Motowidlo, 1999). In empirischen Studien zeigte sich, dass das Verhaltensbeschreibungs-Interview zu leicht höheren prädiktiven Validitäten führt als das situative Interview (z. B. Campion, Campion & Hudson, 1994; Krajewski, Goffin, McCarthy, Rothstein & Johnston, 2006; Pulakos & Schmitt, 1995). Motowidlo (1999, S. 187) erklärt dies damit, dass die Interviewten bei Fragen über vergangenes Verhalten von habituellen Verhaltensmustern berichten: „The more consistent the behaviors are across situations in the past, and the more similar the situations are to situations that occur frequently on the job, the more valid those questions should be." Zukunftsorientierte Fragen erfassen dabei eher Berufswissen.

Wie ich in Kapitel 2.2 dargestellt habe, nimmt man auch bei den SJTs eine ähnliche Unterscheidung vor, indem der Testbearbeiter entweder die Frage nach der Verhaltenstendenz (*would do*) oder nach dem Wissen (s*hould do*) zu beantworten hat. Hier zeigt sich, dass die would do-Instruktion dazu führt, dass die SJT-Werte höher mit Persönlichkeitsskalen korrelieren, wohingegen die should do-Instruktion zu höheren Korrelationen mit Intelligenzmassen führt (McDaniel, Hartman, Whetzel & Grubb, 2007). Dabei beziehen sich die Verhaltenstendenz-Instruktionen eher auf schon erlebtes Verhalten („Wie haben Sie sich in der Vergangenheit typischerweise in einer solchen Situation verhalten?"), die Wissens-Instruktionen eher auf hypothetisches Verhalten („Was sollte in dieser Situation getan werden?"). Somit spricht vieles dafür, bei der Konstruktion eines SJTs zur Erfassung von Persönlichkeitsmerkmalen Situationen zu wählen, welche einen hohen Bekanntheitsgrad haben und bei der Instruktion nach in der Vergangenheit gezeigtem Verhalten zu fragen.

Wählt man den Bekanntheitsgrad als Auswahlkriterium der Acts, schränkt dies jedoch die Breite an unterschiedlichen Situationen deutlich ein, da man hochprototypische aber exotische Situationen ausschliessen muss. Wie in Kapitel 7.3 erwähnt, befassen sich Wissenschaftler schon seit 50 Jahren mit dem Zusammenhang der Breite von in der Personalselektion verwendeten Konstrukten und der Validität der jeweiligen Testverfahren, dem *bandwidth-fidelity dilemma* (Cronbach & Gleser, 1965). Aktuellste Studien und Meta-Analysen (Bergner, Neubauer & Kreuzthaler, 2010; Dudley, Orvis, Lebiecki & Cortina, 2006; Hough & Oswald, 2008; Rothstein & Goffin, 2006) zeigen auf, dass mit der Verwendung von Persönlichkeitsskalen, welche eng gefasste Konstrukte messen, eine bessere Vorhersage der Berufsleistung möglich ist, als mit der Verwendung von globalen Persönlichkeitsdimensionen. Dass sich jedoch sowohl mit eng als auch mit breit gefassten Persönlichkeitskonstrukten Arbeitsleistung vorhersagen lässt (z. B.

Barrick & Mount, 2003; Rothstein & Jelly, 2003; Warr, Bartram & Martin, 2005), kommentieren Rothstein und Goffin (2006) wie folgt:

> Generally, broader criterion measures may likely fit broader personality measures, although the magnitude of the correlation will likely be low. More specific criteria may be a better fit with narrow personality traits and the magnitude of the correlation can be expected to be larger. However, if there is a sound theoretical or conceptual case for expecting any particular personality construct to be related to a particular performance criterion measure, this would be more important than how broad or narrow the personality measure or criterion is. (S. 164)

Grundsätzlich kann also davon ausgegangen werden, dass eine Verschmälerung der im Fragebogen gemessenen Konstrukte eher positive Auswirkungen auf die prädiktive Validität haben wird. Voraussetzung dafür ist jedoch, dass die gewählten Konstrukte *überhaupt* relevant für die zu erbringende Leistung sind. Dies sollte eigentlich der Fall sein, wenn es sich um Dimensionen handelt, welche anhand einer seriös durchgeführten Anforderungsanalyse bestimmt wurden. Trotzdem kommt man nicht umhin, eine empirische Bestätigung für diesen Zusammenhang zu liefern.

Im Zusammenhang mit der Berücksichtigung des Bekanntheitsgrades der Item-Stämme lässt sich auch die Hypothese formulieren, dass sich durch eine weniger spezifische Schilderung der Situationen die Itemkennwerte verbessern, da sich die Testbearbeiter mehr auf in der Vergangenheit gezeigtes Verhalten beziehen können, was zu konsistenterem Antwortverhalten führen könnte. Hingegen könnte die weniger spezifisch und eindeutig geschilderte Ausgangslage auch dazu führen, dass die Testbearbeiter mehr Freiheiten bei der Interpretation der Situation haben, was dann bezogen auf die zu erfassenden Dimension zu heterogenerem Antwortverhalten führt. Um die hier aufgeworfenen Fragen zu klären, müssten entsprechend konzipierte Laborstudien durchgeführt werden.

Zur Akzeptanz des Leadership-Fragebogens

Eines der Ziele dieser Testkonstruktion war, dass die Stellungspflichtigen den Leadership-Fragebogen gut akzeptieren. Die beiden Labor-Studien zeigten auch eindeutig auf, dass der Leadership-Fragebogen besser akzeptiert wird als ein herkömmlicher Persönlichkeits-Fragebogen. Ein Problem stellt jedoch die Augenscheinvalidität des Leadership-Fragebogens dar, welche die Probanden der Akzeptanz-Studie I deutlich am tiefsten von allen Akzeptanz-Aspekten einstuften. Dies ist nachvollziehbar, da sich die Items zur Augenscheinvalidität auf die Ver-

gleichbarkeit mit dem Berufsleben beziehen („Die Aussagen widerspiegeln Anforderungen, die auch im Berufsleben von einer Führungsperson gefordert werden." resp. „Die Aussagen spiegeln Anforderungen wider, die auch im Militärdienst gefordert sind."), die Leadership-Items jedoch mehrheitlich Situationen aus dem Privat-Alltag betreffen. Um eine höhere Augenscheinvalidität zu erzielen, müssten die Items somit Situationen aus dem Alltag von unterem militärischen Kader enthalten. Hierbei ergibt sich jedoch das Problem, dass die Stellungspflichtigen diese Situationen nicht aus eigener Erfahrung kennen und sich gewisse unter ihnen daran stören, dass sie einen Test bearbeiten müssen, welcher Situationen aus dem Militärdienst enthält. Eventuell liesse sich die Augenscheinvalidität auch verbessern, indem Situationen aus dem Alltagsleben gewählt werden, welche einen starken Bezug zu den erfolgsrelevanten Situationen im Militär aufweisen. Da es sich dabei aber um Führungssituationen handeln müsste, stellt sich das Problem, dass die meisten 19jährigen noch nie eine Führungsposition bekleidet haben und so auch über keine entsprechenden Erfahrungen verfügen. Wie ich zudem weiter oben schon ausgeführt habe, erzielen hypothetische Angaben zu in der Zukunft stattfindendem Verhalten weniger gute Validitätswerte als Aussagen, welche sich auf konkret gezeigtes Verhalten in der Vergangenheit beziehen. Als einzig valable Möglichkeit bleibt, den Stellungspflichtigen den Zusammenhang zwischen den im Leadership-Fragebogen dargestellten Situationen und den Anforderungen an militärisches Kader im Rahmen der Testinstruktionen zu erklären. Dass solche Informationen zu den eingesetzten Testverfahren dazu führen können, dass die Testbearbeiter diese besser akzeptieren, konnte in Studien nachgewiesen werden (Burns, Siers & Christiansen, 2008; Noon, 2006; Truxillo, Bauer, Campion & Paronto, 2002). Wie ich in Kapitel 5.4 zu den zehn Regeln des Gilliland-Modells ausführlich beschrieben habe, stellt der Tätigkeitsbezug einen der wichtigsten Aspekte für das Empfinden von Fairness in einem Selektionsprozess dar. „Once again, improving face validity increases perceived justice" (Jones, 1991).

Zu leicht anderen Ergebnissen führte die Befragung der Stellungspflichtigen mit dem Akzept!-Fragebogen zu den drei in den Rekrutierungszentren eingesetzten Testverfahren Leadership-Fragebogen, Persönlichkeits-Fragebogen und Intelligenztest: Die Einschätzung des Leadership-Fragebogens unterscheidet sich kaum von derjenigen des Persönlichkeits-Fragebogens. Es scheint so zu sein, dass der Leadership-Fragebogen nur in einem direkten Vergleich deutlich besser bewertet wird als ein herkömmlicher Persönlichkeits-Fragebogen. Immerhin beurteilten die Stellungspflichtigen den Leadership-Fragebogen deutlich besser als den Intelligenztest – ein Hinweis auf das in der Literatur beschriebene *justice dilemma* (Cropanzano, 1994), welches besagt, dass Bewerber bei in der Perso-

nalselektion eingesetzten Verfahren diejenigen mit einer hohen prädiktiven Validität tendenziell als unfair einstufen.

Die beiden schrittweisen Regressionsanalysen der Akzeptanzdimensionen auf das Erleben der Fragebogenbearbeitung respektive auf die Gesamtbeurteilung des Leadership-Fragebogens ergaben, dass das Layout respektive die Darstellung den wichtigsten Prädiktor darstellt. Weitere wichtige Prädiktoren sind die Alltäglichkeit der geschilderten Situationen, die Augenscheinvalidität/Messqualität und das sich Hineinversetzen-Können in die Situationen. Ob die in meiner Studie befragten jugendlichen Probanden und Stellungspflichtigen die Testung als interessant und angenehm erleben, hängt für sie somit stark mit einem sie ansprechenden Layout zusammen. Damit hat sich meine Vermutung bestätigt, dass sich eine die Situation illustrierende Fotografie auf die Akzeptanz des Leadership-Fragebogens auswirkt. Der zweite wichtige Aspekt für das Erleben und die Gesamtbeurteilung des Testverfahrens sind die Augenscheinvalidität und die Bekanntheit der geschilderten Situation – einer der in der Akzeptanzforschung am besten gesicherte Befunde (z. B. Ryan & Ployhart, 2000). Da unsere Probanden und die Stellungspflichtigen jedoch genau diesen Aspekt beim Leadership-Fragebogen deutlich tiefer als die meisten anderen Akzeptanzmerkmale eingestuft haben, zeichnet sich nun noch deutlicher ab, dass Anstrengungen unternommen werden müssen, um die Augenscheinvalidität zu erhöhen. Diesbezüglich schneiden nach der herkömmlichen Vorgehensweise entwickelte SJTs deutlich besser ab: Sie haben erwiesenermassen einen deutlich höheren Tätigkeitsbezug als Intelligenztests und Persönlichkeits-Fragebogen (Van Vianen, Taris, Scholten & Schinkel, 2004).

Zum Zusammenhang des Leadership-Fragebogens mit anderen Persönlichkeitsmerkmalen und Intelligenz

Das primäre Ziel dieser Arbeit war die Entwicklung eines als Situational Judgment Test konzipierten Fragebogens, welcher die a priori definierten Persönlichkeitsdimensionen Durchsetzungsfähigkeit, Kontaktfähigkeit und Verantwortungsbewusstsein erfasst. Damit bin ich der Forderung von McDaniel und Nguyen (2001, S. 109) nachgekommen: „New technologies need to be developed for better-targeted situational judgment tests to assess constructs of interest." Zur Überprüfung, ob ich dieses Ziel auch erreicht habe, verglich ich den Leadership-Fragebogen mit verschiedenen Leistungstests und Persönlichkeitsskalen. Dabei zeigt sich, dass die drei Skalen des Leadership-Fragebogens wie erwartet Konstrukte erfassten, welche in keinem Zusammenhang mit kognitiven Leistungsmassen – allgemeine Intelligenz, Textverständnis und Merkfähigkeit – stehen

(Korrelationen zwischen -.09 und .08). Wie in Kapitel 2.3 beschrieben, weist die Meta-Analyse über 79 Einzelstudien von McDaniel, Morgeson, Finnegan, Campion und Braverman (2001) eine Korrelation von SJTs mit Intelligenz von ρ = .46 auf. Somit kann als erwiesen gelten, dass die von mir gewählte Konstruktionsmethode zu einem Instrument führt, welches weder direkt noch indirekt kognitive Leistungsfähigkeit erfasst.

Es bleibt die Frage, ob die Skalen des Leadership-Fragebogens auch die intendierten Persönlichkeitseigenschaften abbilden. Dies konnte ich unter anderem mit dem Vergleich der Skalen des Leadership-Fragebogens mit dem NEO-PI-R nachweisen. Vor allem die Dimensionen Durchsetzungsfähigkeit und Kontaktfähigkeit korrelieren mittel bis hoch mit den zu erwartenden Big Five-Persönlichkeitsfaktoren. Bei der Skala Verantwortungsbewusstsein gestaltet sich die Interpretation der Korrelationen zu den Big Five schwieriger, dies aber auch schlicht aus dem Grund, weil diese Eigenschaft als solche nicht direkt im NEO-PI-R abgebildet wird.

8.4 Bedeutung der Ergebnisse und weiterführende Studien

Mit dieser Testkonstruktion konnte ich aufzeigen, dass es möglich ist, mit einer entsprechenden Vorgehensweise einen SJT zu entwickeln, welcher a priori definierte Persönlichkeitsdimensionen erfasst. Zudem liefere ich einen eindrücklichen Beleg für die Nützlichkeit des Wertequadrates bei der Erfassung der Persönlichkeit. Meines Wissens gibt es noch keine Studie, welche belegt, dass sich die in einem Persönlichkeits-Inventar umgesetzten vier Wertequadranten empirisch bestätigen lassen. Weiter konnte ich anhand der Regressionsanalysen die – zumindest für jugendliche Testbearbeiter – herausragende Wichtigkeit des Layouts eines Testverfahrens für dessen Gesamtbeurteilung nachweisen, welche sogar die der Augenscheinvalidität übertrifft. Auch wenn diese Erkenntnisse für die Konstruktion von Persönlichkeits-Fragebogen keineswegs revolutionär sind, zeigen sie doch alternative Möglichkeiten auf und geben Anregungen und Hinweise für die Entwicklung von Verfahren für spezifische Einsatzzwecke oder Bewerbergruppen.

Auf Grund der gemachten Erfahrungen und der gefundenen Resultate drängen sich folgende weiterführende Arbeiten und Studien auf:

Anhand der Analyse des Antwortverhaltens bei der Dimension Durchsetzungsfähigkeit – zum Beispiel mittels der Methode des *loud thinkings* oder anhand eines kognitiven Pretests (Prüfer & Rexroth, 2000) – sind Anhaltspunkte für die ungenügende Reliabilität dieser Skala zu finden. Vor allem ist die Hypothese zu klären, ob es den Testbearbeitern nicht eindeutig klar ist, welche der aufgeführten vier Verhaltensalternativen die im Hinblick auf eine Kaderposition beste ist.

Soll die Akzeptanz der in den Rekrutierungszentren eingesetzten Testverfahren weiterhin überprüft werden, so muss der Akzept!-Fragebogen für diesen Einsatzzweck angepasst werden. Dessen testpsychologische Überprüfung ergab, dass sich dieser nicht für den Einsatz in den Rekrutierungszentren eignet, da einzelne Fragen für leseschwache Stellungspflichtigen nicht verständlich sind und sich die darin enthaltenen Skalen faktorenanalytisch nicht abbilden lassen: So können die Stellungspflichtigen zum Beispiel nicht zwischen der Messqualität und der Augenscheinvalidität eines Testverfahrens unterscheiden. Anhand bestehender Messinstrumente ist zu entscheiden, welche Aspekte der Akzeptanz für den spezifischen Einsatz der Testverfahren in der Rekrutierung von Bedeutung sind. Dabei gilt es zu beachten, dass sich die gewählten Aspekte gut voneinander unterscheiden lassen und dass die Aussagen einfach und leicht verständlich formuliert sind.

In einem Laborexperiment ist zu untersuchen, welche Punkte in der Testinstruktion zu erwähnen sind – zum Beispiel der Einsatzzweck des Testverfahrens oder der Grund, weshalb Alltagssituationen ausgewählt wurden –, damit sich die Augenscheinvalidität des Leadership-Fragebogen erhöht. Die erfolgversprechendste davon ist anschliessend im Realsetting in der Rekrutierung zu überprüfen.

8.5 Literaturverzeichnis

Angleitner, A., & Demtröder, A. I. (1988). Acts and dispositions: A reconsideration of the Act Frequency Approach. *European Journal of Personality, 2*, 121–141.

Annen, H. (2000). *Förderwirksame Beurteilung. Aktionsforschung in der* Schweizer Armee. Frauenfeld: Huber.

Barrick, M. R., & Mount, M. K. (2003). Impact of meta-analysis methods on understanding personality–performance relations. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 197–222). Mahwah, NJ: Erlbaum.

Bergner, S., Neubauer, A. C., & Kreuzthaler, A. (2010). Broad and narrow personality traits for predicting managerial success. *European Journal of Work and Organizational Psychology, 19*, 177–199.

Block, J. (1989). Critique of the Act Frequency Approach to personality. *Journal of Personality and Social Psychology, 56*, 234–245.

Burns, G. N., Siers, B. P., & Christiansen, N. D. (2008). Effects of providing pretest information and preparation materials on applicant reactions to selection procedures. *International Journal of Selection and Assessment, 16*, 73–77.

Buss, D. M., & Craik, K. H. (1980). The frequency concept of disposition: Dominance and prototypically dominant acts. *Journal of Personality, 48,* 379–392.

Buss, D. M., & Craik, K. H. (1984). Acts, dispositions, and personality. In B. A. Maher & W. B. Maher (Eds.), *Progress in experimental personality research* (Vol. 13, pp. 242–301). New York, NY: Academic Press.

Campion, M. A., Campion, J. E., & Hudson, J. P. (1994). Structured interviewing: A note on incremental validity and alternative question types. *Journal of Applied Psychology, 79*, 998–1002.

Cinite, I., Duxbury, L. E., & Higgins, C. (2009). Measurement of perceived organizational readiness for change in the public sector. *British Journal of Management, 20*, 265–277.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological Tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Cropanzano, R. (1994). The justice dilemma in employee selection: Some reflections on the trade-offs between fairness and validity. *The Industrial–Organizational Psychologist, 31*, 90–93.

Dudley, N. M., Orvis, K. A., Lebiecki, J. E., & Cortina, J. M. (2006). A meta-analytic investigation of conscientiousness in the prediction of job performance: Examining the intercorrelations and the incremental validity of narrow traits. *Journal of Applied Psychology, 91*, 40–57.

Evers, A. (2001). The revised Dutch rating system for test quality. *International Journal of Testing, 1*, 155–182.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327–358.

Gloor, A. (1993). *Die AC-Methode. Assessment Center. Führungskräfte beurteilen und fördern.* Zürich: Orell Füssli.

Green, S. B., & Stutzman, T. (1986). An evaluation of methods to select respondents to structured job-analysis questionnaires. *Personnel Psychology, 39*, 543–564.

Green, S. B., & Veres, J. G. (1990). Evaluation of an index to detect inaccurate respondents to a task analysis inventory. *Journal of Business and Psychology, 5*, 47–61.

Helwig, P. (1948). Das Wertequadrat. *Psyche, 2 (1)*, 121–127.

Helwig, P. (1967). *Charakterologie*. Freiburg: Herder.

Hoenle, S. (1996). *Führungskultur in der Schweizer Armee.* Frauenfeld: Huber.

Hough, L. M., & Oswald, F. L. (2008). Personality testing and industrial-organizational psychology: Reflections, progress, and prospects. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 272–290.

Hughes, G. L., & Prien, E. P. (1989). Evaluation of task and job skill linkage judgments used to develop test specifications. *Personnel Psychology, 42*, 283–292.

Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. *Journal of Applied Psychology, 67*, 577–580.

Janz, T. (1989). The patterned behavior description interview: The best prophet of the future is the past. In R. W. Eder & G. R. Ferris (Eds.), *The employment interview: Theory, research, and practice* (pp. 158–168).

Newbury Park, CA: Sage.

Jetter, W. (2008). *Effiziente Personalauswahl. Durch strukturierte Einstellungsgespräche die richtigen Mitarbeiter finden* (3., aktual., überarb. und erw. Auflg.). Stuttgart: Schäffer-Poeschel.

Jones, J. W. (1991). Assessing privacy invasiveness of psychological test items: Job relevant versus clinical measures of integrity. *Journal of Business and Psychology, 5*, 531–535.

Kersting, M. (2008). Zur Akzeptanz von Intelligenz- und Leistungstests. *Report Psychologie, 33*, 420–433.

Krajewski, H. T., Goffin, R. D., McCarthy, J. M., Rothstein, M. G., & Johnston, N. (2006). Comparing the validity of structured interviews for managerial-level employees: Should we look to the past or focus on the future? *Journal of Occupational and Organizational Psychology, 79*, 411–432.

Krüger, C. & Amelang, M. (1995). Bereitschaft zu riskantem Verhalten als Trait-Konstrukt und Test-Konzept: Zur Entwicklung eines Fragebogens auf der Basis des Handlungs-Häufigkeits-Ansatzes. *Diagnostica, 41*, 35–52.

Larsen, R. J., & Buss, D. M. (2001). *Personality psychology. Domains of knowledge about human nature.* Boston, MA: McGraw-Hill.

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 65*, 422–427.

Latham, G. P., & Wexley, K. N. (1982). *Increasing productivity through performance appraisal.* Reading, MA: Addison-Wesley.

Lindley, P., Bartram, D., & Kennedy, N. (2008). *EFPA review model for the description and evaluation of psychological tests. Test review form and notes for reviewers* (Version 3.42). European Federation of Psychologists' Associations. Heruntergeladen am 1. Juli 2010 von www.efpa.eu/download/9044bd41c7953b956876e06c797f8c9f

Maurer, T. J., & Tross, S. A. (2000). SME committee vs. field job analysis rating: Convergence, cautions, and a call. *Journal of Business and Psychology, 14*, 489–499.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L., III. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60,* 63–91.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman,

E. P. (2001). Predicting job performance using situational judgment tests: A clarification of the literature. *Journal of Applied Psychology, 86,* 730–740.

McDaniel, M. A., & Nguyen, N. T. (2001). Situational judgment tests: A review of practice and constructs assessed. *International Journal of Selection and Assessment, 9,* 103–113.

McHenry, J. J., & Schmitt, N. (1994). Multimedia testing. In M. G. Rumsey, C. B. Walker, & J. H. Harris (Eds.), *Personnel selection and classification* (pp. 192–232). Hillsdale, NJ: Erlbaum.

Moser, K. (1989). The Act-Frequency Approach: *A conceptual critique. Personality and Social Psychology Bulletin, 15*, 73–83.

Motowidlow, S. J. (1999). Asking about past behavior versus hypothetical behavior. In R. W. Eder & M. M. Harris (Eds.), *The employment interview handbook* (pp. 179–190). Tousand Oaks, CA: Sage.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology, 75,* 640–647.

Muck, P. M., Höft, S., Hell, B. & Schuler, H. (2006). Die Konstruktion eines berufsbezogenen Persönlichkeitsfragebogens. Integration von Interpersonalem Circumplex, Fünf-Faktoren-Modell und Act Frequency Approach. *Diagnostica, 52*, 76–87.

Mummendey, H. D. & Grau, I. (2008). *Die Fragebogen-Methode* (5. überarb. u. erw. Auflg.). Göttingen: Hogrefe.

Noon, A. L. (2006). *Job applicants' testing and organizational perceptions: The effects of test information and attitude strength.* Unpublished doctoral dissertation, University of Nebraska – Lincoln.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Opgenoorth, W. P. (1979). Die Messung der Anforderungen einer Führungsperson. *Personalführung, 11*, 221–223.

Ostendorf, F. & Angleitner, A. (2004). *NEO-PI-R. NEO-Persönlichkeitsinventar nach Costa und McCrae. Revidierte Fassung.* Göttingen: Hogrefe.

Osterlind, S. J. (1998). *Constructing test items: Multiple-Choice, constructed-response, performance, and other formats* (2nd ed.). Boston, MA: Kluwer Academic Publishers.

Prüfer, P. & Rexroth, M. (2000). Zwei-Phasen-Pretesting. In P. P. Mohler & P. Lüttinger (Hrsg.), *Querschnitt. Festschrift für Max Kaase* (S. 203–219). Mannheim: ZUMA.

Pulakos, E. D., & Schmitt, N. (1995). Experience-based and situational interview questions: Studies of validity. *Personnel Psychology, 48*, 289–308.

Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review, 16*, 155–180.

Rothstein, M. G., & Jelly, R. B. (2003). The challenge of aggregating studies of personality. In K. R. Murphy (Ed.), *Validity generalization: A critical review* (pp. 223–262). Mahwah, NJ: Erlbaum.

Ryan, A. M., & Ployhart, R. E. (2000). Applicants' perceptions of selection procedures and decisions: A critical review and agenda for the future. *Journal of Management, 26*, 565–606.

Schulz von Thun, F. (1989). *Miteinander Reden 2. Stile, Werte und Persönlichkeitsentwicklung.* Reinbek bei Hamburg: Rowohlt.

Tannenbaum, R. J., & Wesley, S. (1993). Agreement between committee-based and field-based job analyses: A study in the context of licensure testing. *Journal of Applied Psychology, 78*, 975–980.

Truxillo, D. M., Bauer, T. N., Campion, M. A., & Paronto, M. E. (2002). Selection fairness information and applicant reactions: A longitudinal field study. *Journal of Applied Psychology, 87,* 1020–1031.

Van Vianen, A. E. M., Taris, R., Scholten, E., & Schinkel, S. (2004). Perceived fairness in personnel selection: Determinants and outcomes in different stages of the assessment procedure. *International Journal of Selection and Assessment, 12*, 149–159.

Warr, P., Bartram, D., & Martin, T. (2005). Personality and sales performance: Situational variation and interactions between traits. *International Journal of Selection and Assessment, 13*, 87–91.

Westermann, F. (Hrsg). (2007). *Entwicklungsquadrat. Theoretische Fundierung und praktische Anwendungen.* Göttingen: Hogrefe.

Wilson, M. A., Harvey, R. J., & Macy, B. A. (1990). Repeating items to estimate the test-retest reliability of task inventory ratings. *Journal of Applied Psychology, 75*, 158–163.

Anhang 8.1 Aufstellung der in dieser Arbeit verwendeten Stichproben

	Jahr	Teilnehmer	Stichprobengrösse
Erstellung des Anforderungsprofils			
Erstellung der Anforderungsprofile (Interviews)	2009	Berufsmilitärs	22
Gewichtung der Verhaltensweisen des Anforderungsprofils	2009	Berufsmilitärs	60
	2010	Gruppenführer Zugführer	55 51
Generierung des Ausgangsmaterials für die Itemkonstruktion			
Generierung der Acts	2002	Schüler	38
Einstufung der Prototypizität	2002	Bekannte / Unteroffiziers-Schüler	37
Testentwicklung			
Testentwicklung (4 Studien)	2001 – 2003	Rekruten / Unteroffiziere	977
Testüberprüfungsstudie I (Kürzung von 13 auf 10 Items pro Skala)	2003	Stellungspflichtige	7'871
Testüberprüfung und -weiterentwicklung			
Testüberprüfungsstudie II	2008	Stellungspflichtige	19'801
Pretest mit Likert-Skalierung	2007	Stellungspflichtige	1'017
Vergleichsstudie Antwortformate F-C - likert	2009	Studierende	100
Validierungsstudie Persönlichkeits-Vergleichsfragebogen	2003	Unteroffiziere	442
Validierungsstudie Leadership - Intelligenz	2005	Stellungspflichtige	F-C: 17'040 likert: 930
Validierungsstudie Leadership - Persönlichkeit	2005	Stellungspflichtige	F-C: 16'994 likert: 930
Validierungsstudie Leadership - NEO-PI-R	2009	Studierende	100
Akzeptanzuntersuchungen			
Bekanntheitsgrad der Situationen	2009	Studierende	35
Akzeptanz des Leadership-Fragebogens I	2002	Berufsschüler, Gymnasiasten, Rekruten	131
Akzeptanz des Leadership-Fragebogens II	2002/03	Unteroffiziers-Schüler	690
Akzeptanz des Leadership-Fragebogens III	2008/10	Stellungspflichtige	606

Anmerkung. Die Angaben zur Stichprobengrösse beziehen sich auf die nach der Bereinigung des Datensatzes für die Berechnungen einbezogenen Probanden.

Lebenslauf

Patrick Boss

*11.02.1969

Sigriswil / BE

1988	Maturität (Realgymnasium) an der Kantonsschule Baden
1989 – 1996	Studium der Angewandte Psychologie, Psychopathologie des Kindes- und Jugendalters und Neurophysiologie an der Universität Zürich
1996 – 2004	Assistent an der Abteilung Angewandte Psychologie am Psychologischen Institut der Universität Zürich (Prof. Dr. F. Stoll)
1999 – 2011	Projektleiter „Entwicklung und Umsetzung psychologischer Testverfahren für die Rekrutierung A XXI"
2004 – 2011	Projektleiter an der Fachrichtung Sozial- und Wirtschaftspsychologie am Psychologischen Institut der Universität Zürich (Prof. Dr. K. Jonas)
2004 – 2007	Chefpsychologe der Rekrutierung der Schweizer Armee (Milizfunktion)
seit 2011	Wissenschaftlicher Mitarbeiter und Berater am Zentrum für Verkehrs- und Sicherheitspsychologie am Institut für Angewandte Psychologie der Zürcher Hochschule für Angewandte Wissenschaften